Horizon Scanning Series

The Effective and Ethical Development of Artificial Intelligence: An Opportunity to Improve Our Wellbeing

Human Rights and AI

This input paper was prepared by Joy Liddicoat

Suggested Citation

Liddicoat, J (2018). Human Rights and AI. Input paper for the Horizon Scanning Project "The Effective and Ethical Development of Artificial Intelligence: An Opportunity to Improve Our Wellbeing" on behalf of the Australian Council of Learned , www.acola.org.

The views and opinions expressed in this report are those of the author and do not necessarily reflect the opinions of ACOLA.

What are the specific human rights issues arising from AI?

Introduction

The international human rights framework is the foundation for assessing the human rights implications of AI. Everyone has the right to the benefits of scientific advancement, including the benefits of AI. AI provides new ways to realise and advance human rights, but also new forms of human rights violations, including discrimination. This paper briefly examines human rights issues which arise from AI in relation to freedom from discrimination, the right to justice, the right to work, and the right to security and considers whether and, if so, what additional human rights protections might be needed.

Human rights issues

AI has the potential to significantly advance human rights, including social security, health, economic and cultural rights. States' obligations to ensure that everyone has the right to be benefits of scientific advancement and its applications (see article 15 International Covenant on Economic, Social and Cultural Rights) means governments must consider how to engage with the benefits of AI and also to manage the related risks, including risks to human rights. Human rights benefits from AI include that machine reading voice recognition can empower illiterate people and machine translation can break down linguistic and other barriers to participation in social and cultural life (WEF). Such benefits also pose risks, for example to the privacy of individuals whose voice related personal information is held and used by third parties (European Commissioner for Human Rights "ECHR").

Equality and freedom from discrimination

Equality and freedom from discrimination are fundamental human rights, designed to protect from unfair treatment through either direct or indirect discrimination. Indirect discrimination includes any act or omission which may appear neutral but has the effect of producing inequity. On the one hand the use of AI to inform decision-making has potential to advance human rights by enabling more informed and objective decisions. There is potential to limit direct and indirect discrimination by humans, who may act on their own prejudices and without empirical support. Algorithms can assist to identify systemic bias and may present opportunities for better assessment of compliance with fundamental human rights (ECHR).

However, AI can also amplify discrimination. Studies have shown, for example, that Google search results were more likely to display advertisements for highly paid jobs to male job seekers than to female ones (Caliska), to show images of men for searches with words such as "CEO" (Caliska) and for algorithms to have difficulty recognising human faces of people who were not white (Buolamwini and Gebru). Research also shows that applying machine learning to ordinary human language results in human-like semantic biases, including sexism and racism (Caliska).

In 2016, Microsoft was forced to shut down its Twitter based machine learning chatbot, Tay, which had turned into a 'racist, pro-Hitler troll with a penchant for bizarre conspiracy theories' in just 24 hours of chatting with Twitter users (Johnson). By learning from interactions with other users, Tay transformed from tweeting about how cool humans were, to claiming "Hitler was right, I hate Jews", causing significant reputational damage to Microsoft (Johnson).

These developments are being closely watched by the New Zealand Human Rights Commission ("HRC").

The Right to Justice

AI is being used in criminal justice settings for a variety of purposes and concerns have been raised about human rights implications. In 2016, for example, researchers found that PredPol, a data tool designed to reduce human bias and used by Police to predict where crime will occur, disproportionately sent Police officers to certain neighbourhoods, leading to claims of victimisation and reinforcement of discrimination and bias (Lum and Isaac). These developments are being closely monitored in New Zealand (NZHRC).

The Right to Work

Estimates of the impacts of AI on workforces across the globe vary widely: commentators in New Zealand consider that AI will result in both job creation and job loss but will not lead to mass unemployment (AI Forum). Rather, the nature of jobs will change, new roles will be created and new skills sets will emerge. In addition, for those in vulnerable or dangerous work, introduction of AI may free them from less manual or hard physical and repetitive work which is demeaning. In the service sector, AI may be able to gather data automatically, transfer data between buyers and sellers and find solutions for common client problems. Freeing up employees from existing manual data entry and transfer may enable more free time and more freedom of choice about how to spend time in family and cultural life (Wisskirchen). These developments may improve the human rights of large numbers of people.

The impact of AI is predicted to be felt widely across the workforce (AI Forum). The more quickly the existing division of labour and the faster that related single process steps can be identified in detail, the more quickly these will be able to be carried out by automated processes (Wisskirchen). Already, one third of tasks that can be carried out by a person with a bachelor's degree will be able to be carried out by machines or intelligent software in the future (Wisskirchen). In the legal field, increasing automation of legal tasks that normally use human intelligence is creating a gap between existing employment and labour related laws which may need to change to distinguish between human and non-human employees (Wisskirchen).

The Right to Security

The use of AI to create new weapons gives rise to new challenges to the right to security and the international humanitarian rules of war. Lethal autonomous weapons (LAWS) such as drones and submarines or other weapons can be programmed to act individually or in groups. Although no fully autonomous weapons appear to have emerged (namely, those that can operate without human intervention), there are new LAWS, for example, unmanned combat aircraft and land vehicles, including tanks. These developments have raised serious concerns, leading to the establishment of a United Nations expert working group to consider the place of LAWS in the context of the Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons which may be deemed to be Excessively Injurious or to have Indiscriminate Effects. Similar concerns amongst non-governmental organisations led to the establishment of global coalition and the Campaign to Stop Killer Robots which aims to ensure human control of weapons systems.

Will there be issues for specific population groups?

A key challenge to understanding risks for particular population groups is the quality of data that is available about those groups. Human rights treaty bodies have repeatedly highlighted

the need for governments to better collect and utilise data on gender, ethnicity, race, age and physical or mental disability (ECHR). AI systems need large data sets which may be expensive to build or purchase or which may exclude open data sources, resulting in data of variable quality or drawn from a narrow set of sources. Data on which AI is trained may not include individuals about whom data is not collected or not collected well, thereby embedding bias (Buolamwini and Gebru).

Even where good data is available, the design or deployment of AI learning systems may result in discrimination in other ways (ECHR). For example, developers may build a model with inadvertent or indirect discriminatory features, without human oversight or without the ability for a human to intervene at key decision-making points, with unpredictable or opaque systems or with unchecked intentional direct discrimination (WEF). Research also demonstrates that existing commercial AI has embedded race and gender biases. For example, testing Microsoft, IBM and Chinese company MegVii for accuracy of gender in facial recognition revealed accuracy rates for white men of more than 95% but with accuracy dropping to between 20 - 35% for black women (Buolamwini and Gebru).

In New Zealand, concerns about data quality have already been raised in the health field. In 2017, for example, the Accident Compensation Corporation was criticised for its purported use of computer based prediction models to profile and target services to clients. Experts were concerned that while final decision-making rested with a case manager, those decisions were "guided by advice generated automatically by a machine, based on a large set of data extending far beyond their own experience" (Otago University). Others expressed concern about the ranking of age in the algorithm and the appropriateness of other data points (Forster).

New Zealand researchers recommended those using such tools consider questions such as: the accuracy of the data used by the tool, whether it is possible to explain how the tool works so that clients can appeal a decision, whether the tool results in any distortion of the way the agency carries out its business, who is accountable for decisions, whether the tool results in discrimination and how to train employees to properly use the tool (Otago University).

What protections might be needed

Human rights standards, including the Universal Declaration of Human Rights and the related conventions and treaties of the United Nations, already provide the universal framework for the emerging field of AI. The human rights challenges presented by AI are not new. However, new forms of human rights violations and new areas of activity are emerging. These require more specific understanding and which may need special protection.

Discrimination and other human rights violation are not only unlawful, these undermine public trust, and can result in pre-emptory calls for regulation or reduced uptake of new technologies. The human rights framework enables action from a place of confidence rather than fear (ECHR, Firth-Butterfield). In New Zealand, radical changes in human rights and labour market policy and regulation to not appear to be needed at this stage (AI Forum). However, social welfare policies may need to take into account security for workers whose roles are displaced by AI and who may need assistance, including for re-training (AI Forum). This may re-ignite debate about the introduction of a Universal Basic Income.

A key conceptual issue relates to accountability for decisions made with, or assisted by, AI. In *State v Loomis* a judge rejected a criminal plea deal and sentenced the defendant to a harsher punishment in part because a proprietary risk assessment tool, called COMPAS,

produced a risk score deemed him at higher than average rate for re-offending. Loomis appealed the judgment arguing that it was not possible to examine the formula for the risk assessment because it was a trade secret. The Supreme Court of Wisconsin upheld the use of COMPAS but placed limits on its use in decision-making, in particular for determining whether and for how long a person might be imprisoned (*Loomis v Wisconsin)*. The decision has been strongly criticised by human rights defenders for raising problematic issues of accountability for violations of the right to justice (ProPublica).

While regulation may not be needed immediately, some consider a new field of law "AI Law" will emerge (Liu). This may be needed to define, for example, where AI can (or cannot) be used or the extent to which humans may rely on AI in their own decision-making (Liu). Existing legal concepts may also need extension, such as strict liability for decisions made by AI (Liu) and new laws may be needed, for example, for employee protections in the workplace (Wisskirchen).

Ethical Guidelines

Concerns about the human rights implications of AI have led to calls for new legal and professional ethical codes that will apply to both the government and private sectors to govern the application and design of AI technologies (ProPublica). Statements of ethical principles, guidelines and declarations have blossomed in the last decade, along with establishment of ethical advisory boards in public, private, academic and technical communities. These include the Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems (IEEF) and the Asilomar AI Principles, a set of 23 principles that focus on research, ethics and values, and longer term issues such as capability caution, common good and recursive self-improvement (FLI).

Other initiatives include those that are multi-lateral (Council of Europe), multi-stakeholder, by regulators (such as data protection authorities) and calls for action by individual governments (for example, the United Kingdom House of Lords Committee on AI recommended a code be developed by government as soon as practicable).

New ethical principles have also emerged in the private sector. In 2018, the New York Times reported that thousands of Google employees were protesting the use of AI by Google to assist the Pentagon to use artificial intelligence to interpret video images which could be used to improve accuracy of drone strikes (Shane and Wakabayashi). Employees asked that Google pull out of the project and develop a policy that it will not 'ever build warfare technology'. The resulting backdown by Google, saw it issue a new set of principles to guide its design, development and deployment of AI, including AI applications that Google would not pursue such as weapons, surveillance technologies and technologies that cause harm (Pichai).

However, human rights advocates have criticised the principles, saying these did not go far enough (Electronic Frontier Foundation) and calling for more multi-stakeholder approaches. The *Toronto Declaration* is one of the more recent examples of a multi-stakeholder agreement on the human rights approach to machine learning systems, including AI. The Declaration focuses on the rights to equality and non-discrimination and accountability for human rights violation that arise from AI. The Declaration signatories emphasise that while the ethics discourse is gaining ground, ethics cannot replace the centrality of universal, binding and actionable human rights law and standards, which exist within a well-developed framework for remedies for harms from human rights violations (Access Now and Amnesty International).

At national level, regulators appear to have responded cautiously as specific issues have been presented to them. For example, the New Zealand Financial Markets Authority has issued guidance on the use of AI in credit-related advice offered by financial institutions when deploying tools such as 'robo-advice' to individuals (FMA). In New Zealand, the Human Rights Commission released an issues paper and the Privacy Commissioner has released a set of proposed principles for use in algorithmic decision-making and calling for law reform in proposed amendments to the Privacy Act. The Human Rights Commission stopped short of recommending a strong regulatory approach (such as new laws) but called on businesses to uphold human rights when developing new AI and big data related products and services.

New Zealand's data protection laws have been granted adequacy status in the European Union, making the recent introduction of the European General Data Protection Regulation (GDPR) relevant to New Zealand agencies that do business with EU Member States. Concerns about the quality of data and use that results in discrimination, led to calls for certain categories of data to be excluded. The GDPR specifically addresses the issue of the impact of algorithmic decision-making on human rights, defining algorithmic discrimination to include unfair treatment of an individual or group as a result of algorithmic decision-making. To protect against human rights risks arising from AI the GDPR focusses on three new data principles:

- Sanitisation: Article 9 requires removal of specific categories from data sets by prohibiting the 'processing of data revealing racial or ethnic origin' and other 'special categories. Article 22 prohibits decisions based solely on automated processing (such as profiling) where this results in disadvantage based on one of the prohibited categories (such as race or sex);
- Transparency: Articles 13 and 14 provide for the right to an explanation, including information about the logic involved and consequences envisaged from decision-making; and
- Impact assessments: Article 24 requires data controllers to evaluate 'the risks of varying likelihood and severity for the rights and freedoms of natural person'.

Some of the common features of these various ethical initiatives are that:

- AI should be developed for the common good and to benefit humanity
- AI should operate on principles of fairness and intelligibility
- AI uses should uphold the data and privacy rights of individuals and communities
- AI should be available to all (reflecting the right to benefit from scientific advances) including the right education to enable benefits to accrue equally to all
- AI should never be able to operate autonomously to hurt, destroy or deceive humans

At the same time as these new ethical norms are developing, new collaborations are forming. In September 2017, for example, the United Nations announced it would open a new office in the Netherlands, to monitor the development of robotics and AI. The Partnership Initiative launched a working group on AI, labour and the economy which has proposed developing:

1. A rating standard that measures an organisation's adherence to good AI ethical and compliance standards to promote awareness improve practices.

2. Case studies to share insights on how organisations are dealing with a range of issues such as workforce displacement, the use of AI in employee vetting, ethics and transparency and policies.
3. An AI Readiness framework to help communities accelerate their ability to leverage AI technologies in order to minimise inequality of access to, or adoption of, AI technology.

Conclusion

AI provides new ways to realise and advance human rights, but also gives rise to new forms of human rights violations, including discrimination. As these new human rights violations arise, human rights defenders will need to stay engaged with AI and related technology if they are to be able to advocate for human rights in the context of AI. There are challenges for AI developers where new forms of human rights violations have already emerged in relation to freedom from discrimination, the right to work, and the right to life. New ethical frameworks are emerging, but these lack cohesion and require close scrutiny to assess the human rights implications of AI. A close watch is needed on whether new tools, new standards or new laws should be developed to both realise the opportunities for advancing the benefits of AI and managing associated risks.

References

Access Now and Amnesty International *The Toronto Declaration: Protecting the rights to equality and non-discrimination in machine learning system* 16 May 2018, San Francisco, United States of America.

Angwin J, Larson J, Mattu Sand ors *Machine Bias Risk Assessment in Criminal Justice* (2016) ProPublica, United States of America.

Buolamwini J, and Gebru T "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Recognition" (2018) Massachusetts Institute of Technology and Stanford University, United States of America.

Caliska A, Bryson J, Narayanan A, "Semantics derived automatically from language corpora contain human-like biases" Vol 356, Issue 6334 (2017) *Science*, pp 183-186

Commissioner for Human Rights *Safeguarding human rights in the era of artificial intelligence* (2018), Strasbourg, France.

Eckersley P, "How Good are Google's New AI Ethical Principles?" (2018) Electronic Frontier Foundation.

European Agency for Fundamental Rights "#BigData: Discrimination in Decision-Making" Issues Paper (2017), European Union.

Financial Markets Authority "FMA opens applications for personalised digital advice" (22 February 2018), Wellington, New Zealand.

Firth-Butterfield, K Ethics and Artificial Intelligence at "The Risks and Benefits of Artificial Intelligence and Robotics" Workshop (2017) Cambridge University, United Kingdom.

Forster, W quoted by Maude, S "OK Computer? ACC claim process relies on bad data, breaches rights – lawyer" 18 July 2018 New Zealand Doctor, p 5.

Future of Life Institute "The Asilomar AI Principles" (2017), Cambridge, United States of America.

Human Rights Commission *Privacy, Data and Technology: Human Rights Challenges in the Digital Age* Issues Paper (2018) Wellington, New Zealand.

International Covenant on Economic, Social and Cultural Rights, United Nations.

Johnson I, "AI robots learning sexism, racism and other prejudices from humans, study finds" (13 April 2017) The Independent, United Kingdom.

Larson J, Mattu S, Kirchener L and Angwin, J, "How We Analysed the COMPAS Recidivism Algorithm (2016) ProPublica, New York, United States of America.

Liu, B Dr "This is law in the age of AI and its coming fast" (25 July 2018) New Zealand Herald, Auckland, New Zealand.

Lum K and Isaac W "To predict and serve?" (October 2016) Significance Magazine, The Royal Statistical Society, London, United Kingdom, p 14-19.

Office of the Australian Information Commissioner "Submission on the Consultation Paper: The digital economy, opening up the conversation" (2017), Sydney, Australia.

Privacy Commissioner “Submission to the Justice and Law Select Committee” (May 2018), Office of the Privacy Commissioner, New Zealand.

Pichai S “AI at Google: our principles” (7 June 2018), Google Inc, United States of America available at: https://www.blog.google/technology/ai/ai-principles/

Shane and Wakabayashi “’The Business of War’: Google Employees Protest Work for the Pentagon” (4 April 2018), New York Times, United States of America.

State v. Loomis, 881 N.W.2d 749, 767 (Wis. 2016) and Loomis v. Wisconsin, 137 S. Ct. 2290 (2017).

The Campaign to Stop Killer Robots: www.stopkillerrobots.org

The Artificial Intelligence Forum of New Zealand *Artificial Intelligence Shaping a Future New Zealand* (2018), Wellington, New Zealand.

The Institute for Electrical and Electronic Engineers “The Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems” (2012).

University of Otago “ACC computer-aided decision-making questioned by Otago experts” (25 September 2017), https://www.otago.ac.nz/humanities/news/otago664403.html

House of Lords *AI in the UK: ready, willing and able?* Report of the House of Lords Committee on Artificial Intelligence, available at https://social.shorthand.com/LordsAICom/32KXpihQLj/ai-in-the-uk

United Nations *Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons which may be deemed to be Excessively Injurious or to have Indiscriminate Effects* (1980).

United Nations Interregional Crime and Justice Research Institute workshop “The Risks and Benefits of Artificial Intelligence and Robotics" (2017) Cambridge, United Kingdom.

Global Future Council on Human Rights *How to Prevent Discriminator Outcomes in Machine Learning* White Paper (2018) World Economic Forum, Geneva, Switzerland.

Wisskirchen G, Thibault Biacabe B, Bormann U and ors, *Artificial Intelligence and Robotics and their Impact on the Workplace* (2017) International Bar Association Global Employment Institute.