THE HUMANITY IN ARTIFICIAL INTELLIGENCE

Can the ‘perfect’ algorithm have a moral conscience?

May 2019

Briefing Paper

TRPC

Contents

Introduction

The title is intentionally misleading and suggests that artificial intelligence (AI) reflects human values and morals. It is instructive to think of intelligent machines as capable of processing information from their environment as having two dimensions to their activity beyond that of operating or functioning according to their design, such as to issue alerts and warnings. First, what they do, their actions, as a result of machine learning is consistent with the way they were designed to operate even though their environment may have changed beyond the original assumptions made by their design. Second, the 'decisions' they make, their behaviour, do not contradict those that would be made by humans. This could of course include the decision to kill because humans who are accountable to norms and rules of society also make such decisions, and the grey areas of human decision-making inevitably will become the grey areas of AI behaviour. Water-boarding, for example, is widely regarded by society as a form of torture, yet elements within the security forces and their political supporters regard it as a legitimate way to gain information that could save lives – although the evidence seems to suggest otherwise.

The important point is that society in some way or another – laws, through majority political decisions, etc. – makes the choice of actions based upon a sense of human morality. Morality is defined in dictionaries as a set of inner principles that guide human sensibilities.[1] Explicit guidance on how actions should reflect that morality is often stated in laws, rules, injunctions, etc., as a set of ethics. Ethics is therefore defined, and differentiated from morals, as a set of external rules governing actions or behaviour. However although ethics may be a universally accepted and desirable concept, there is no universal agreement on what may or not be considered ethical. Again, taking an extreme example, the death penalty is seen in some societies or by some persons as ethnical and in others not, but the cases in which the issue comes to be debated derive from similar underlying moral concerns.

Ethics is often motivated by moral behaviour and plays a pivotal role in human interaction as it sets boundaries that are universally accepted and are often adhered to.[2] How then can this be applied to AIs and their interactions with human beings? This poses a further dilemma if AI is being developed in different societies even where they share similar moral principles. Can there be a set of ethical rules that is universally-accepted? In the extreme cases, such as causing death, is the practical answer that the ultimate decision must be taken by a human who can be held responsible and accountable? Or can those responsible for the design and operation of the machine be held responsible assuming they can be identified and are still living? Mostly the dilemmas are not as extreme as involving life and death, but nevertheless can impact upon individuals and humanity in severe ways.

Issues such as misidentification and bias in decision-making are already well-documented. In May 2019 California decided to ban the use of public facial recognition technologies, even by law enforcement agencies, as a reaction against mistakes, yet by contrast China is developing the world's most sophisticated facial recognition applications using the vast amounts of data a population of over 1.3 billion people can provide to perfect the performance of AI. Does perfection overcome the problem of mistakes, only to shift the danger to human decisions about the direction of AI? And finally, what is the

[1] It is worth noting that morality tends to be common across persons of both religiosity and those of no religion, suggesting that it is innate to the welfare of societies even when what is considered to be an action that properly reflect that morality differs according to time, place and circumstances.

[2] University of Texas, https://ethicsunwrapped.utexas.edu/glossary/morals

likelihood that the science fiction of today of AI machines making *all* the relevant decisions without recourse to human intervention – indeed by-passing human intervention – becoming a real future threat? Before that moment is reached, a practical perspective needs to bear in mind two points: first, AI is here and is working and can bring many benefits, including the elimination of mistakes made by humans, and second, progress towards a universally-accepted set of ethics for AI use is required as we challenge AIs to become better decision-makers than ourselves.

While a sense of moral direction may be innately built into human beings, it is unlikely the same can be said for AI. Despite being programmed to possess cerebral functions that allow an AI to respond independently, the limit and extent of these reactions are often fixed within set parameters. In addition, the moral and ethical 'consciousness' is to be instilled within an AI's algorithm, it is likely that these would follow the moral compass of its programmer.

What are we trying to regulate?

Apart from the programmer's biasness, the behaviour of an AI is also a reflection of the cultural nuances and biases within existing data-sets. Given much of human history is tainted with biases on social classes, gender, race, etc, we may yet be unable to produce the balanced data-sets needed to avoid prejudice.[3] This could lead to potentially serious consequences if AIs make automatic decisions based on bias data. And even when AIs are fed "new data", AIs are unlikely to be able to differentiate between legitimate, and illegitimate data. As an example, in 2016 Microsoft launched its experimental AI chatbot "Tay", which was designed to mimic language patterns. After being online for only 16 hours, the chatbot was forced to be shut down after publishing racist and sexist tweets.[4] Tay's public interactions grew to be more immoral when users realised that the algorithm was able to mimic their behaviour and respond similarly. Therefore, while the algorithm was a variable in which the programmers could control, the data-set (in the form of interactions from other Twitter users) was an unexpected component. The system also did not have in place 'moral safety measures' that would have allowed it to discern between right and wrong. And while critics were quick to point out the limitations in Microsoft's AI chatbot, the algorithms weren't so much to blame as the 'bad' information fed to Tay by human users.

While the use and application of AI has seemingly limitless potential, AI's inability to distinguish between social cues and inappropriate actions, makes the need for a self-initiated check and balance within the AI's algorithm all the more pertinent. With the growing prevalence of AI in our day to day lives, the consequences of incorrect predictions or any missteps in social cues potentially result in significant and undesirable consequences.

The crux of the issue is thus how do we ensure that 'ethical' safeguards keep pace with AI development so that innovation remains responsible and trustworthy. In this space there are often many other issues worthy of regulation such as personal privacy, fairness, accountability, etc. The World Economic Forum (WEF) in 2016 listed out its "Top 9 ethical issues in artificial intelligence".[5] The list remains relevant today and worth thinking about. The Top 9 ethical issues in AI according to the WEF are:

[3] Medium, https://medium.com/applied-innovation-exchange/the-good-the-bad-and-the-ugly-of-artificial-intelligence-and-machine-learning-3f7e663c317a

[4] The Verge, https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist

[5] WEF, https://www.weforum.org/agenda/2016/10/top-10-ethical-issues-in-artificial-intelligence/

1. **Unemployment**: the automation of jobs is likely to fill in gaps for activities that can be automated, and while some natural displacement is expected this also leads to the creation of other forms of jobs as previous experiences with the Industrial revolution and digital revolutions have shown. Importantly, if automation is able to free up more time for more personal endeavours could that then lead to a higher quality of living? Shorter working hours so more time can be spent with family, learning, or do structural shifts in the economy negate this as utopian.

2. **Inequality**: While the use of AI could help reduce human labour, does that also mean a reduction of wages paid out by companies who employ AI rather than humans? This could threaten to further increase in inequality. Alternatively, can we use AIs to reduce the inequality gap. According to an Oxfam report, in 2019 the top 26 richest people in the world were equally as rich as 50% of the world's poorest. With the use of AI expected to lead to cost savings and more efficient wealth generation, how do we more equitably distribute the wealth created by machines?[6]

3. **Humanity**: as AIs are getting better at replicating human behaviour, such as in the use of chat bots, how then does that affect the way we interact with (a) machines, and (b) humans. Are AIs becoming more "human", or are we becoming less "human"? Further, where clickbait articles and video game rewards published by AI encourage us to perform certain actions, could this be used to nudge society towards either more beneficial or greater behaviour?

4. **Artificial stupidity:** could AI's ability to learn be used against it. For example, by being fed false information, fake news to produce wrong predictions and misdirect users? As the Microsoft Tay example demonstrates, there are a lot of deviant material and people in the world who could try and exploit and abuse AI. In such cases, where such events occur, who should be held accountable and responsible?

5. **Racist Robots:** removing all biases from robots is a difficult task. Robots and AI algorithms are built by human beings who have implicit biases, on a subconscious or conscious level. While an AI algorithm can be designed to be completely agnostic to race, gender, religion, and orientation, minimising bias will also involve active learning (on the robot's part) as well as unbiased data set. How is that active learning guided?

6. **Security:** innovation is always accompanied by the potential for harm, as with every new technology, and there are bound to be weak points in the system that will make it susceptible to cyberattacks and/or misuse. How do we then ensure that cybersecurity keeps pace with AI development.

7. **Evil genies**: as with every development, there can be unintended consequences that we may not have considered. In the case of AI, even though there may not be any evil intentions, a lack of understanding of the full context may lead to unintended consequences.

[6] The Guardian, https://www.theguardian.com/business/2019/jan/21/world-26-richest-people-own-as-much-as-poorest-50-per-cent-oxfam-report

8. **Singularity:** more commonly attributed to science fiction works, where AIs become smarter than humans – predictions on when this could happen range between 5-30 years. How then can we ensure humans stay in control of such complex intelligence system? And is that desirable?

9. **Robot rights**: currently, robots are machines equipped with AI algorithms and bear more similarities to a car or toaster than to any biological beings. However, in the conceivable future robots could potentially develop conscious thoughts and feelings, and should they then be given adequate rights on their own?[7]

How are governments and industry tackling AI ethics?

Given these complexities and challenges, governments and industry all around the world are wrestling with frameworks and the formation of advisory committees and ethical boards to tackle these issues. Accenture for example, has suggested a framework for a government implementing a responsible AI and Robotics (Ethical Framework):[8]

1. Set up an AI advisory body
2. Gather intelligence on and participate actively in the development of such codes internationally
3. Develop core ethical principles
4. Encourage the development of sector specific codes

While many governments have already begun this approach and are at steps 1 and 2, step 3 continues to remain a work in progress. Such principles are currently used more as guidelines, for fears any premature regulation which inhibit the development and innovation of AI use. To better understand how these guidelines apply to AI development and application, governments in Europe and Singapore are inviting organizations to 'pilot' test these frameworks to better understand how to work in the real-world, and where further refinements are necessary.

The following section documents examples on how governments and industry are approaching the complexities of AI ethics development.

Government

Many governments around the world have developed a set of ethical guidelines for R&D of AI in their respective counties. One of the first in Asia was South Korea that spells out the rights and responsibilities of manufacturers, owners and of the robots themselves, even according them fundamental rights "to exist without fear of injury or death" and "to live an existence free from systematic abuse." This has not stopped protests over the design of robots as weapons by academic institutions for the military. Japan has similarly been very actively promoting the industrial development and use of AI and robotics and in 2017 the "Artificial Intelligence Technology Strategy Council" published the "Artificial Intelligence Technology Strategy" which includes a section on the, 'Principle of ethics (respect human dignity and individual autonomy)'.

In Oceania, the New Zealand government has also released its Government Algorithm Transparency Report providing a review of the government's use of algorithms through fourteen self-assessed government agencies on their respective use of their algorithms.[50] One of the recommendations made is

[7] Discover, http://blogs.discovermagazine.com/crux/2017/12/05/human-rights-robots/#.XO0BmtMzbOQ
[8] Accenture, https://www.accenture.com/gb-en/company-responsible-ai-robotics

for published information to better explain how algorithms may inform decisions affecting ordinary people. In Australia, the Department of Industry, Innovation and Science released a discussion paper that detailed the government's framework towards approaching AI ethics. Developed by CSIRO's Data61, the consultation ran for two months in 2019 and sought opinions from both the private sector and the public.[9] This initiative follows the Federal Government's announcement in its 2018 budget statement that a funding would be developed specifically for a national AI ethics framework.[10] The paper also emphasises that the framework would tweak existing laws and ethical principles so they can be applied in the context of AI technologies, as opposed to rewriting new laws or ethical standards.

Likewise In China, the development of AI is considered a national priority to place China as a global industrial leader, but as an MIT paper phrases it "the Chinese government sees AI as a tool for social governance" such as the social credit system, which inevitably leads to criticism from the Western perspective. But there is a drive to introduce ethical considerations, particularly coming from leading Chinese companies very aware of public concerns over privacy, leading to the setting up in January 2019 of a Chinese Association for Artificial Intelligence.

In 2019, Singapore announced its guidelines on AI ethics at the World Economic Forum at Davos as " the first in Asia to provide detailed and readily implementable guidance to private sector organisations using AI." The guidelines reflect several AI initiatives "including an Advisory Council on the Ethical Use of AI and Data chaired by former attorney-general V.K. Rajah." Where ethics and industrial R&D meet for governments is at the intersection of standards, guidelines and frameworks. Singapore's Model AI Governance Framework includes provisions for human-centricity, transparency, fairness and explainability and comprises guidance and measures promoting the responsible use of AI that should be adopted by organisations.

In the US, the 2016 'National Artificial Intelligence Research and Development Strategic Plan'[11] highlights AI technology as transformative and pivotal in furthering the national industrial, economic and social priorities of the country, with a recognition of the potential dangers and the need for ethical guidelines. But somewhat predictably, the 'American AI Initiative' launched in February 2019 made no mention of ethics.[12]

In contrast, in April 2019 the EU published a set of ethical guidelines[13] under seven headings: Human agency and oversight, Technical robustness and safety, Privacy and data governance, Transparency, Diversity, non-discrimination, and fairness, Environmental and societal well-being, and Accountability. At the heart of the EC's AI Guidelines, is how AI applications are based on fundamental rights as enshrined in the Charter of Fundamental Rights of the European Union (EU Charter), and in relevant international human rights law including respect for human dignity, freedom of the individual, respect for democracy, justice and the rule of law, equality, non-discrimination and solidarity, and citizen's rights.

[9] Australian Government, Department of Industry, Innovation and Science, https://consult.industry.gov.au/strategic-policy/artificial-intelligence-ethics-framework/

[10] Australian Computer Society, https://ia.acs.org.au/article/2018/budget-2018--ai-boost-with-an-ethical-focus.html

[11] The Networking and Information Technology Research and Development Program, https://www.nitrd.gov/news/national_ai_rd_strategic_plan.aspx

[12] Future of Life Institute, https://futureoflife.org/ai-policy-united-states/

[13] The Verge, https://www.theverge.com/2019/4/8/18300149/eu-artificial-intelligence-ai-ethical-guidelines-recommendations

Mostly recently the OECD Council adopted the "OECD Principles on Artificial Intelligence" on 22 May 2019, representing the first intergovernmental standard to be implementable.[14] In addition to the 36 OECD member countries, Argentina, Brazil, Colombia, Costa Rica, Peru, and Romania, have expressed and signed their support. The four key principles in the standard are: i) inclusive growth, sustainable development and well-being, ii) human-centred values and fairness, iii) transparency and explainability, and iv) robustness, security, and safety.[15]

Industry

Among the best known industry initiatives to address AI and ethics is the 'Partnership on AI' (PAI) founded by Amazon, Facebook, Google, DeepMind, Microsoft, IBM, and later Apple. PAI was established in 2016 to advance "public understanding in the sector, setting societal and ethical best practices for AI research on 'Fair, Transparent, and Accountable AI' through an open platform for discussion and engagement, and to produce standards for future researchers."[16]

However, IT companies are unavoidably conflicted between commercial imperatives, technology-driven innovation and social responsibilities. After just one week Google closed down its own 'Advance Technology External Advisory Council' (ATEAC) following a controversy over appointments to its advisory board. Behind ATEAC was the idea of addressing ethical issues around AI and other emerging technologies, including concerns of racial bias. Likewise Amazon came under fire for misrepresented technical aspects of research, which had suggested that Amazon's facial recognition tool, Rekognition, was less accurate towards women and people of colour.[17] Shareholders however, rejected the banning any sale of facial recognition software to law enforcements.[18] The tech firm went on to release a statement saying that it had not received any reports of law enforcement clients abusing the tool. Amazon also pushed the ball back to lawmakers, stating that the responsibility and decision on restrictions should lie with the individual companies who have decided to use the service.[19]

Elsewhere, the Japanese Society for Artificial Intelligence (JSAI)'s AI Ethics Committee published its own set of "AI Society Ethics Guideline" in 2017.[20] Comprised of nine key points, the guidelines place equally weighted importance on the respect for the diversity of culture, honesty with regard to the technical limitations of any AI technology, and to accommodate the "variety of voices" within Japanese society.[21] Current members of the JSAI include NTT Com, the Cyber Agent Corporation, the National Research and Development Corporation New Energy and Industrial Technology Development Organisation, as well as the Ministry of Internal Affairs and Communications (MIC), and the Information and Communications Policy Research Institute.[22]

[14] Epic.Org, https://epic.org/2019/05/oecd-announces-ai-principles-4.html

[15] OECD Legal Instruments, https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

[16] Partnership on AI, https://www.partnershiponai.org/

[17] QZ, https://qz.com/1586544/ai-experts-want-amazon-to-stop-selling-facial-recognition-to-law-enforcement/

[18] USA Today, https://www.usatoday.com/story/tech/2019/05/22/amazon-facial-recognition-shareholders-reject-banning-software-sales/3770927002/

[19] BBC, https://www.bbc.com/news/technology-48422321

[20] Japanese Society for Artificial Intelligence Ethics Committee, http://ai-elsi.org/archives/471

[21] Japanese Society for Artificial Intelligence Ethics Committee, http://ai-elsi.org/wp-content/uploads/2017/02/%E4%BA%BA%E5%B7%A5%E7%9F%A5%E8%83%BD%E5%AD%A6%E4%BC%9A%E5%80%AB%E7%90%86%E6%8C%87%E9%87%9D.pdf

[22] Japanese Society for Artificial Intelligence, https://www.ai-gakkai.or.jp/about/sponsors/

China’s Tencent’s vision of ethics in AI or “tech for good” is slated to be shaped closely by the company’s belief that the idea of good should be dependent and enforced by the individual, rather than being enforced by the government. Equally, this notion can be conceived as being in harmony with the Communist Party’s concern to portray its governance as promoting the public good.

Uncannily similar to Google’s founding mantra “do no evil”, some employees, however, have criticised Tencent’s board as lacking teeth, as the tech giant continues to cooperate with the Chinese government which is known to ‘request’ for data from organizations.[23] At the time of writing, official publications documenting the Chinese company’s agenda for AI were yet to be published.[24]

Conclusion/Way Forward

The issues on how to ensure algorithms are programmed without bias and how to hold AI accountable if something goes wrong continues to be the challenge that companies and governments are facing when handling data and the algorithms that feed on it. AI Ethics and Bias are two sides of the same coin that continue to need further study and discussion.

Actionable ethical principles will also need to take into consideration the different uses of AI in different contexts. For example, the ethical issues that could arise from the use of AI in autonomous weapons, will be extremely different from the use of AI in autonomous vehicles. In addition, ethical issues involved in the use of AI for employment screening in Asia is likely to be different from the use of AI for employment screening in America.

At the same time, the lack of transparency, poor accountability, unfairness, and bias comes with the territory of there being millions of lines of code in each application. This makes it difficult to ascertain the exact point in the coding where the values are inculcated and a decision is reached. The sheer complexity of the coding and range of AI functionality makes its especially important that an ethical approach is designed into the system prior to its implementation, especially when the context for its future use is hard to predict. As mentioned, some of the ways algorithmic bias can unwittingly occur is when the data used to train an AI model is tainted by factors such as existing human biases, incomplete knowledge, misrepresented data, or the ongoing interactions of users.

Organisations can address these ethical issues through augmentation, by placing people at the centre, and augmenting the workforce by applying the capabilities of machines so people can focus on higher-value analysis, decision-making and innovation. The AI algorithms should also be able to provide clear explanations for the actions they take as a form of communication and collaboration. Further, to develop a universal, clear, explicit, and transparent code of ethics, a universally recognised governing body may also need to be established to coordinate and align – which is no easy task.

Concluding with more questions than answers to think about: Are commonly shared concerns purely motivated by the need for trust (and the lack of it)? Are the developers and users of AI or the people who regulate and set the rules universally trusted? Can AI itself be programmed to evaluate the necessary levels of trust that are required?

[23] The Wall Street Journal, https://www.wsj.com/articles/why-china-is-keeping-a-close-eye-on-tech-giant-tencent-1531047602

[24] Financial Times, https://www.ft.com/content/f92abc38-6bb8-11e9-80c7-60ee53e6681d

Additional Resources

1. Personal Data Protection Commission of Singapore - A Proposed Model AI Governance Framework, https://www.pdpc.gov.sg/Resources/Model-AI-Gov

2. European Commission - Ethics guidelines for trustworthy AI, https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

3. CMS - Artificial Intelligence CMS Insights section, https://cms.law/en/GBR/Insight/Artificial-Intelligence

4. MIT – Moral Machines, http://moralmachine.mit.edu/

5. Organisation for Economic Co-operation and Development (OECD) - Principles on Artificial Intelligence, https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449