From ethical principle to responsible action in the age of artificial intelligence: a short paper to spark discussion

The World Economic Forum has referred to artificial intelligence as the engine that drives the Fourth Industrial Revolution [1]. McKinsey and Company call it the "transformational technology of our digital age" [2] and note that deep learning techniques alone have the capacity to deliver "between $3.5 trillion and $5.8 trillion in value annually" [2]. Accenture research on the impact of AI in twelve developed economies notes that AI could double annual economic growth rates in 2035 [3] and AlphaBeta flags that digital data innovation already accounts for 11% of GDP in advanced economies [4]. In short, artificial intelligence is transforming our economy in genuinely massive ways.

But there are growing signs that our voracious pursuit of that economic transformation is arriving without a full consideration of the potential societal transformation and impact that accompanies it. Recently, artificial intelligence systems deployed by the criminal justice system have shown racial biases [5], IBM Watson produced "unsafe and incorrect" cancer treatment recommendations [6], and hundreds of thousands of British students were exposed to the socio-economic biases of an algorithmic grading system which downgraded marks for students from poorer areas [7]. Moreover, "government institutions are increasingly deploying automated decision making systems… that are [often] untested and poorly designed for their tasks… Worse, when they make errors and bad decisions, the ability to question, contest, and remedy these is often difficult or impossible" [8].

The emergence of worrying case-studies reflecting rapid development and largely unchecked integration of artificial intelligence systems across sectors has led to a burgeoning collection of ethical frameworks being developed by academic, industry and government parties. The European Commission, CBA, DeepMind, Microsoft, the Partnership on AI, the Future of Humanity Institute, IEEE, the Future of Life Institute, the House of Lords, and the Australian Department of Industry, Innovation and Science (amongst *many* others [9, 10]) have all weighed-in on the defining principles of a more responsible approach to artificial intelligence that has an eye firmly fixed on the public good and prevention of societal and individual harms. The principles, while diverse in description, are largely concordant in direction – appealing to concepts of fairness, accountability, explicability, reliability, and promotion of human wellbeing, rights and agency.

While the emergence of frameworks that promote and endorse these principles is critical, they are largely meaningless without tools, systems and processes that allow for the robust design, evaluation and operation of artificial intelligences against those principles and within those frameworks. Equally, developers and innovators looking to do the right thing need technological, regulatory and legal support to reduce the complexity of baking ethics into artificial intelligence design and development from the start.

We are witnessing a collision of rapidly evolving and partially aligned guidelines, materialising risk that has resulted in real harms to real people, an uncertain regulatory environment, and genuine opportunity for artificial intelligence to deliver transformative value. We cannot wait for the dust to settle. A failure to make best efforts today while waiting for the certainty of tomorrow will only result in the promulgation of the status quo in the field of artificial intelligence. A status quo that has led to remarkable progress, but at the possible cost of our human values.

This brief paper provides an initial sketch of how we may move from ethical principle to responsible action in the very near term. The intent is not to provide a prescription, but to spark a discussion on how we – as a community – should best move forward from where we are today. A forum is scheduled to further the discussion on responsible artificial intelligence practice on 10 December 2020 and we invite you to participate.
https://staff.uts.edu.au/Pages/StaffNotice.aspx?ct=Important+notices&ItemID=7184

Why is artificial intelligence different?

Contemporary artificial intelligence[1] brings with it a number of factors which complicate the development and delivery of responsible systems.

Scale: The grounding of artificial intelligence in software means that it is often and largely liberated from the resource constraints that otherwise limit the deployment and adoption trajectories of traditional services and technologies. When integrated with inherently scalable cloud computing infrastructure, for example, an artificial intelligence solution can rapidly scale from a few users in one location to hundreds of thousands of users globally, with little in the way of profound additional engineering effort. The consequence is that the impact of artificial intelligence solutions can be far reaching and arrive incredibly quickly. At the best of times, it is difficult to anticipate the societal impacts of any technology even at a small-scale level, but when there is potential for near-simultaneous wide-spread (even global) adoption, the risk becomes front-loaded. Incremental adoption enables lessons to be integrated into solution development; the big-bang of AI means that by the time that lessons arrive, it may be too late.

Opacity: The explosion of deep learning as a viable technology has led to innovations in everything from Go to large-scale facial recognition in public surveillance settings. But the power of deep learning carries with it the cost of complexity – such complexity that many deployed solutions are approaching complete inscrutability – even to the developers who created them. Encased in a black box, contemporary artificial intelligence solutions often (inadvertently, but necessarily) obscure their inner workings in a way that few other technologies in history have. And, ultimately, it is difficult to have trust and confidence in something we do not truly, genuinely, understand.

Data: The capacity of contemporary artificial intelligence systems to harvest incredible volumes of complex interconnected data is heralded as a key differentiator over traditional decision making paradigms – rife as they are with subjectivity and claimed expertise. Leveraging a robust evidence base enables us to capitalise on relationships and patterns in data that are difficult – often impossible – for the human eye to see. But therein lies the rub. As our datasets grow in depth and breadth, we grow less able to understand the fine-grained character of the fuel we are feeding our algorithms. We are asking algorithms to learn from something we ourselves cannot fully grasp and the risk is that in the thickets of data lie biases, errors and bad habits that are being absorbed into our most intelligent artificial systems. Couple this with business models that often exploit personal information (with varying levels of informed consent) and we are putting ourselves at risk of not just making bad decisions, but doing so in a way that erodes personal privacies.

Core principles for more responsible artificial intelligence

The scale, opacity and fueling of artificial intelligence may complicate delivery of ethical systems, but the goals can be plainly stated. The depth and breadth of the dozens and dozens of ethical guidelines that have emerged in the last few years is necessary to underpin regulatory and legislative thinking, but we can move forward quickly with a simple distillation of what defines a responsible artificial intelligence system:

Explicability: Artificial intelligence is complicated. Develop methods and processes for explaining how it is being used, what data fuels it and why decisions are being made. Make artificial intelligence auditable.

Fairness: Do not bias outcomes for or against particular groups of people based primarily on their membership in one of those groups. No matter what the data says.

Reliability: Ensure that artificial intelligence systems perform as designed, including beyond initial deployment and across a wide range of conditions. Working in testing does not guarantee working in practice.

[1] For a broad and simple definition, let's state that an artificial intelligence is a computer system that is capable of performing tasks nearing, at or beyond human skill.

Agency: Produce systems that maximise, rather than minimise, human agency. Give people power to control how their data is used. Give people the capacity to override artificial intelligence decision making. Do not build systems that impinge on our personal freedoms.

Privacy: Do not encroach on people's privacy, neither in the design and development of an artificial intelligence system nor in its ultimate application. Privacy standards, legislation, process and practice is well established – there is nothing special about artificial intelligence which should allow it to circumvent or ignore these.

Accountability: Provide clarity on who assumes responsibility for consequences flowing from the use of each artificial intelligence system at each point in the artificial intelligence lifecycle, from initial testing, through trials and to ongoing operation and use. Accountability can particularly become ambiguous in those cases where artificial intelligence informs – but does not dictate – human decision making, or where humans are expected to have an oversight role (as in early autonomous vehicle trials). Up-front articulation of roles and responsibilities is therefore critical to ensuring risk is properly owned and mitigated.

Preservation of rights: Respect human rights and prevailing laws. This should not need to be stated.

Obstructions to moving from principle to action

Given that the principles themselves are simple, it may seem reasonable to conclude that the response is similarly so. Unfortunately, there are a range of barriers that has slowed a genuine integration of ethical principles into real-world AI practice.

A lack of help: As noted, both the data feeding contemporary artificial intelligence systems and the underlying algorithmic structure of those systems can be profoundly complex. To make systems emerging from such complexity explicable and fair is itself far from trivial. We cannot expect development teams to spontaneously become experts in human-centered explanatory systems nor in the rigorous assessment of bias. As such, a longer-term pathway is in the integration of these concepts into data science training programs that target the development of a *responsible* artificial intelligence workforce, rather than simply an artificial intelligence workforce, and an increasing focus on transdisciplinary connection and capability. But this will take time and action is required today. The fastest path forward is for existing development teams to adopt mature and well-tested libraries and software packages that explicitly target fairness and explanation as standard bolt-on solutions to black box systems. But, while the space is rapidly expanding and significant technology leaders are investing (including IBM [11] and Google [12]), all feel some way from representing the complete solution that teams need. At present, the response feels early and piece-meal, and software practice has not coalesced around a consistent approach to integrating and using existing tools.

Distance: The development of robust responsible artificial intelligence systems requires meaningful collaboration that brings users, developers, ethicists and the community together. Yet, at every turn, distance pervades:

- Developers are not trained to understand the nuances of ethics or the language of ethics, but nor are ethicists conversant in the subtleties of artificial technology developments or practical operation.

- When artificial intelligence is viewed as critical intellectual property – as it often is – there is a reluctance to expose the inner-workings of the black box and to bring third parties into genuine review and refinement processes.

- Users are often viewed as clients, and so there is reluctance to reveal system flaws that may jeopardise contracts or delivery schedules.

- The community is exposed to risk, but communicating that risk jeopardises product uptake and success, and development teams are not trained to engage with broad community stakeholders in any case. The matter is only complicated when artificial intelligence is targeting wide deployment which crosses multiple potential user groups across multiple cultural contexts.

- Language, principles, priorities and business processes vary by sector, making it difficult to share knowledge across distinct projects.

In short, friction to engagement exists everywhere.

Carrots, sticks and boxes: There are clear trends towards regulation of artificial intelligence that will bound ethical performance expectations and processes, but with few exceptions, these have not yet been formally enshrined in legislation or recognised global standards. Organisations will remain reluctant to invest in processes if there are no guarantees that those processes will fit within future regulatory and legislative boxes, particularly when current frameworks do not strongly incentivise responsible behaviour or penalise poor behaviour.

Ambiguity: Ethical decision making has been a hot topic of conversation for a couple of thousand years and it is naïve to think that there are unified views on what constitutes the good and the right. Is an autonomous weapon that reduces the risk faced by one set of combatants, while increasing lethality against another combatant producing benefit? Well, it probably depends on which side of that equation you sit, your views on the role artificial intelligence should have in killing, whether we agree with the context in which it is being deployed, and our beliefs about future application and controls. An absence of clarity can be stupefying or exploited – if there is little agreement around what we should be doing then we may choose to do nothing for fear that any path may be harmful or we may choose to do anything, knowing that creative arguments can almost always be generated in support. We may think we know the difference between a system that helps and harms humanity, but in most cases, the line is less sharply defined than it appears on first glimpse.

Getting moving

In the absence of existing processes, we recommend a simple engagement-focused, collaborative and iterative approach to moving from ethical guidelines to adoption of responsible practice in artificial intelligence that is cognisant of the unique character of artificial intelligence and the obstructions to adopting guidelines in practice[2]. Core aspects should focus on:

Maximising engagement: The process should integrate community representatives, external artificial intelligence experts, ethicists, developers and end-users (collectively, *the stakeholder panel*) in an open and iterative process where the goal is to co-design an optimal solution and to learn together. The process should not be a uni-directional critique or adversarial interaction, but a genuine collaboration and an effort to understand and integrate distinct perspectives in a profoundly complex space.

Alignment with guidelines: Noting the consistency across global ethical AI guidelines, the process should include a deliberate focus on explicability, fairness, reliability, agency, privacy, accountability, and human rights. By centering on these elements, the process will result in solutions that fit with emerging global norms, providing a strong measure of future-proofing against future regulation (which will almost certainly be framed in response to existing guidelines).

Transparency: The process should result in publicly accessible outputs that detail the ethical process engaged, the success metrics agreed, the findings produced and the path forward. It should be formally reviewed ratified by all participants in the ethical engagement process. In the absence of formal standards and certification, such outputs underline that a deliberate and serious approach to the responsible production of artificial intelligence has been undertaken.

Market awareness: The ethical co-design process should be underpinned by robust protections of intellectual property to enable free discussion and technical review. More than this, though, successful application of the process should result in assets that are of genuine and differentiating market value for the artificial intelligence developer. In the absence of established standards, formal certification is perhaps a step too far, but the generation of simple and clear assets that underline adherence to ethical guidelines and proof points for that

[2] We note that the development of robust legislation, clearer definitions of accountability and responsibility, and formalisation of standards may be critical factors for establishing lasting best practice in the development of responsible artificial intelligence. Our focus here is on near-term actions that can be driven by those active in the artificial intelligence space today, and excludes wider developments which could be driven by government and the international standards community.

adherence should be made available to developers to use in marketing materials and in communications efforts.

Being objective when we can be: Experts in the quantitative assessment of responsible AI objectives, including data bias and explicability, should co-design objective targets for success based on an iterative process with the stakeholder panel. These experts should facilitate the use of relevant software tools and libraries to enable the assessment and provide practical advice on how best to improve performance. The goal should be rapid assessment and system refinement, but also knowledge transfer that better equips development teams to integrate objective ethical review into their ongoing development pipelines.

Being subjective when we must: We should acknowledge that the successful implementation of responsible artificial intelligence cannot be completely assessed against objective quantitative standards. Many aspects of risk, wellbeing and agency are arguable and the process should enable that argument. Central to this is to have a genuine conversation with end-users and those most impacted by system deployment, one that removes the distance generated by technical complexity and jargon. With concerns documented, practical mitigations should be agreed and documented with the stakeholder panel, including how risk and protections should be communicated to users and the wider community.

Applicability across the complete life-cycle: Ideally, any responsible artificial intelligence process should operate across the complete system development life-cycle, from initial design through to ongoing post-deployment maintenance and review. At each phase in the life-cycle, a different form and depth of review and co-design is appropriate. At project commencement, a review of risks will reveal guiding project principles, high-level objectives and processes for enabling iterative review and refinement processes. From the point of system testing, where the shape and specific nature of the final product is better defined, deeper consultation and review in keeping with the points discussed throughout this section become more viable and should be pursued. Post deployment, systems and processes should be established to track ongoing performance to ensure that there is not ethical drift over time, as new data and operational conditions come into play.

Avoiding ethical silos: For a subset of artificial intelligence solutions, existing human research, medical research and privacy standards may apply. In these cases, utilisation of existing ethical review processes in these spaces – which are robust and widely recognised – should be integrated into the development of responsible artificial intelligence systems. The key here is *integration*. There should be a two-way flow between existing ethical processes and the artificial intelligence process to ensure alignment, reduce redundancy and improve efficiency.

Being focused, practical and unambiguous: There is a danger that perfection may become the enemy of the good here. A one-size-fits-all process that necessitates a comprehensive review of every ethical risk associated with a broad range of potential deployment futures and uptake trajectories may make review practically impossible or at least profoundly cost prohibitive for smaller artificial intelligence developers. Given this, we recommend bounding the review to specific objectives, agreed and documented in terms of reference that define the role and purpose of the stakeholder panel *for this specific system*. So long as the bounding is clearly stated in engagement, reporting and public assets, this enables targeted reviews of (for example) high risk areas and more incremental development cycles, without sacrificing transparency and public communications. As an example, a specific initial review focused on an artificial intelligence for criminal justice may focus on assessing potential racial and regional biases of specific relevance to the initial deployment region.

Next steps

The above process reflects a promising pathway for the practical realisation of responsible artificial intelligence in today's environment, with its focus on a documented and consultative approach to explicability, fairness, reliability, agency, privacy, accountability and human rights. But this paper primarily exists to start a wider conversation around that assertion, the critical aspects of responsible delivery, and the development of codified, sustainable and repeatable practices for the rapidly evolving artificial intelligence community. A process is commencing that will accelerate and deepen that conversation within and beyond UTS. Interested parties are invited to attend the forum and reach out with comments directly to the UTS Data Science Institute by contacting dsadmin@uts.edu.au.

References

1. World Economic Forum. *Shaping the Future of Technology Governance: Artificial Intelligence and Machine Learning*. 29 August 2019]; Available from: https://www.weforum.org/platforms/shaping-the-future-of-technology-governance-artificial-intelligence-and-machine-learning.
2. Chui, M., et al., *Notes from the AI Frontier*. 2018, McKinsey&Company.
3. Purdy, M. and P. Daugherty, *Why Artificial Intelligence is the Future of Growth*. 2017, Accenture.
4. AlphaBeta, *Digital Innovation: Australia's $315B Opportunity*. 2018, AlphaBeta (commissioned by Data61).
5. Angwin, J., et al., *Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks.*, in *ProPublica*. 2016.
6. Ross, C. and I. Swetlitz. *IBM's Watson supercomputer recommended 'unsafe and incorrect' cancer treatment, internal documents show*. 2018 29 August, 2019]; Available from: https://www.statnews.com/2018/07/25/ibm-watson-recommended-unsafe-incorrect-treatments/.
7. Walsh, B., *How an AI grading system ignited a national controversy in the U.K.*, in *Axios*. 2020: https://www.axios.com/4f728465-a3bf-476b-9127-9df036525c22.html.
8. Whittaker, M., et al., *AI Now Report 2018*. 2018, AI Now Institute, New York University.
9. Jobin, A., M. Lenca, and E. Vayena, *The global landscape of AI ethics guidelines.* Nature: Machine Intelligence, 2019. **1**: p. 389-399.
10. Bird, E., et al., *The ethics of artificial intelligence: Issues and initiatives*. 2020, European Parliament.
11. IBM Research. *AI Fairness 360*. 2020 19/11/2020]; Available from: https://aif360.mybluemix.net/.
12. Google Cloud. *Explainable AI (Beta)*. 2020 19/11/2020; Available from: https://cloud.google.com/explainable-ai.