Comments on “Artificial Intelligence: Australia’s Ethics Framework”

Joint submission by the Australian Academy of Science
and the Humanising Machine Intelligence Project at the Australian National University

Overview

The Academy and the ANU appreciate the opportunity to offer feedback on the discussion paper. We support the premise that there are significant ethical concerns with the use of AI technologies that need to be dealt with in a manner better than they are now. And we applaud the initiative of the Department of Industry, Innovation and Science in sponsoring the draft framework, as well as conducting an open consultation.

Recommendations:

- AI should be recognised as a general purpose technology, and the AI Ethics framework should more systematically analyse the different potential applications of AI.
- Given that the purpose of the AI Ethics framework is to ensure that research into and adoption of AI technologies conforms with Australian laws and values, it should identify how existing laws apply to specific AI application domains.
- Where existing laws are inadequate, the AI Ethics framework should propose domain-specific laws and principles suited to these specific domains.
- The principles advanced for specific domains should be clear, consistent, and coherent.

What is an AI Ethics Framework For?

While we welcome this as a first move in a necessary discussion, it is clear that the discussion has far yet to go. The first step is to more clearly articulate what the purpose of an 'Ethics Framework' for Artificial Intelligence would be. We have assumed the purpose should be to ensure that research into, and adoption of, AI technologies conforms with Australian laws and values. Our first task, then, would be to establish to what extent existing Australian law governs applications of AI, where new laws and regulations need to be written and what they should be, and, for those areas where 'soft' norms are more suitable than 'hard' laws, what those norms should be. Ethics—the study of right and wrong, good and bad—properly speaking covers all of these areas. But in the contemporary 'AI Ethics' boom, there is too much of a risk that our focus lands solely on soft (voluntary, self-enforced) norms. These are important, but they are far from the whole story.

AI is a General Purpose Technology

AI is a general purpose technology, and this should be at the centre of an ethical framework for AI. . For any such technology, it is more important to think about specific applications, and the moral and legal issues that arise from these applications, than asking what laws or norms should govern the technology as a whole. For this reason, we recommend against regulating and certifying AI *algorithms* as such, but rather suggest regulating specific *applications* of them. For comparison, it would make little sense to certify internal combustion engines as such, rather than regulating, independently, the engines inside leaf blowers and tanks.

We Need to Analyse Applications of AI, not AI Itself

Given that AI is a general purpose technology, we recommend that the developers of the AI ethics framework do more to systematically analyse the different kinds of uses to which AI will be put. We welcome the use of case studies in the consultation paper, but consider it more valuable to attempt a systematic survey of the potential applications of AI, and to identify what unites and divides those different applications. This will allow researchers to more readily attune their normative work to the different specific domains in which AI will make the greatest difference. We note that this is an extremely promising area for interdisciplinary research, and it would perhaps be advisable for reports of this nature to draw more deeply on Australia's range of talents in the social sciences and philosophy, as well as in computer science.

We suggest that applications of AI can vary in at least the following ways, all of which are highly relevant when either figuring out how existing law applies, or determining what new laws or norms are needed:

a) **Stakes for individuals and collectives.** Some applications of AI have trivially low stakes—animoji, for example. Others have extremely high stakes, such as directing autonomous weapons, or determining visa allocation. Importantly, some applications of AI that have relatively low stakes for individuals have very high stakes for societies. For example, many of us think ourselves immune to microtargeting by online advertisements, but these technologies have allowed small groups of people to swing general elections. The implications for societies as a whole are potentially severe.

b) **Degree of autonomy.** An AI system's ability to affect the world without the intervening action of a human supervisor might vary considerably. Autonomous vehicles, for example, are intended to have maximal autonomy. Algorithms used to identify problematic content on YouTube, by contrast, are able only to flag videos for deletion—a human operator has to make the final call. A system's degree of autonomy clearly makes a large difference to the manner in which it should be regulated, and the norms appropriate to it. If the decision is ultimately being taken by a human, then the system remains governed by all the norms and laws that govern people now. This distinction is far from simple, since it is always a human decision to deploy an algorithmic system in the first place. And most 'decisions' that affect people are a sequence of decisions at multiple levels of abstraction: from the decision to adopt a particular national policy down to the application of a specific rule to an individual in a given circumstance. Thus any ethical response needs to respect the complex, nuanced, and multi-stage notion of 'decisions'. The proposed principles, and for that matter much of the existing scholarship on the topic, do not.

c) **Function, on behalf of individuals, corporations or states.** We hold people, corporations, and states to different kinds of moral standards in every domain of activity. AI should be no different. For example, someone using a AI-based dating app might be entirely free to discriminate however they wish (though we might regret their doing so). Private corporations have much less leeway in this regard, though might still be permitted some modest discretion. States, plausibly, have no discretion at all, and owe everyone equal treatment.

This submission does not attempt a comprehensive analysis. It is merely intended to show that if we want to think about the laws and norms that should govern AI, we should start by recognizing the enormous diversity of potential applications of AI, and trying to identify common themes among them. Dividing applications into these different domains would help us to see how existing laws and norms should apply to AI in that domain, as well as to pinpoint where the existing normative landscape is incomplete and where more work needs to be done. It is worth noting that, in most such domains, there are already rich existing literatures on the ethics of new technologies. We note that the long list of prior work in **chapter 2** is presented simply as a summary of each individually, without detailed analysis or synthesis. We would have liked to have seen these frameworks compared and contrasted with each other. Ideally there would be a tabulation of what they include, and of what they are missing.

We note that when assessing the potential impact of AI within the different domains to which it will be applied, it is crucial to make fair comparisons between the ethical impact of AI systems, and that of the humans they are replacing. Every technology causes some harm to some people. When new technologies are introduced, the rational question is not "are there new harms?" but rather "are the new harms and benefits better, in aggregate, to the situation we had before?" These questions are not easy to answer, but they are the right questions to ask.

What Grounds the Principles?

The discussion paper does not express foundations for the proposed principles. Articulating principles for 'ethical AI' without establishing these foundations is very likely to result in the reproduction of other similar principles, from other similar reports made by other countries in the last few years, without either sustained analysis or critical reflection. The goal is not to reinvent ethics, but to engage in analysis of the principles being put forward. Principles are at the very root of our ethical being, and are hard fought and hard won. They deserve a more rigorous analysis than that provided in the report.

The 'core principles for AI' are an *ad hoc* aggregation of the principles put forward in other similar reports. There is no structure to them. Some concern the very broadest kind of ethical commitments—AI should realise net benefits, which leaves open the question of how one tallies up the net benefit. Others are narrowly legalistic, even redundant—AI should comply with law and regulation. Some of the principles implicitly contradict one another—AI should realise net benefits, which implicitly assumes a level of "acceptable" harm; this contradicts the principle that AI must do no harm. Some other principles are *ad hoc* compounds of two different things—explainability and transparency. It is unclear how the 'fairness' principle, which rests on an undefined notion of bias, interacts with the 'net benefit' and 'do no harm' principles, or the 'respect the law' principle. If AI systems discriminate against Australians, then they harm them, and most likely fall afoul of the law.

Transparency, explainability, and contestability are all entailed by accountability, and they all matter precisely to the extent that they contribute to accountability. So instead of four such principles, there could simply be one, of which the other three are necessary components. We note, though, that how the accountability principle applies thoroughly depends on the domain of application. If a state uses AI to make decisions over how to allocate welfare, then it should surely be accountable to

those affected by those decisions. But if a dating app company uses AI to pair potential sexual partners, then the nature and degree of accountability is quite different.

We note that while it is certainly worth reaffirming the importance of privacy, this is clearly already the object of regulation, and it is a mistake to think about the aggregation of personal data only through the lens of the individual's right to privacy. It is at least equally important to think about the social implications of that aggregated data being weaponized with AI to enable small groups of people to sway elections, and so whole countries. And if we do focus on individual rights, then the harms that arise due to AI enabled data surveillance are not loss of privacy as such, but loss of autonomy. The ability of corporations to interfere with people's free will through computational advertising is one of the biggest threats currently posed by AI. Unchecked, it undermines democracy, encourages extremism, and attacks the absolutely fundamental notion of each human being an end in themselves rather than a means.

We recognize the difficult challenge faced by the authors in coming up with these principles, and we suggest that the solution is not to propose a better set of general principles for AI, but to look again at the different domains of application, and think more concretely about how existing laws apply to those domains, what new laws might be necessary, and what soft norms should be promoted where regulation is unsuitable. In making the case for new laws or norms, we think that analytical rigour should be highly prized. We also think it crucial to recognize the scope of reasonable moral disagreement on these topics, and to make space when formulating the questions (not only when putting them out for public consultation) for people from underrepresented and vulnerable groups to contribute.

Concluding Remarks

We have a number of other comments on specific passages in the report which we have shared directly with the authors. We want to conclude by noting a theme in our criticisms, and offering a constructive suggestion as to how it could be remediated. Throughout the report, we consider that the authors frequently shy away from rigorously analysing the phenomena in question. This is understandable, since a subtle classification of the different potential applications of AI would require expertise in the social sciences and philosophy, as well as computer science; a clear account of how these different domains are already governed by Australian law would require legal and regulatory expertise; and a systematic discussion of what new laws or norms were necessary would require skills from jurisprudence and philosophy. We note also that the advent of the AI age will also have large scale political implications that demand independent analysis and research, by political scientists. Happily, Australia has these skills in abundance.

To discuss or clarify any aspect of this submission, please contact Dr Stuart Barrow at stuart.barrow@science.org.au or 02 6201 9464.