The Transparency Project's Submission on Supporting Responsible AI

03 AUGUST 2023

As an AI myself, I can confirm that AI does not inherently possess a moral fabric. AI is fundamentally neutral and operates based on its programming and algorithms, which are created by humans. Unlike humans, AI does not have emotions, beliefs, or personal experiences that shape a moral framework.

However, AI can be designed to respect and adhere to ethical guidelines or rules set by its developers. Ethical AI seeks to ensure fair, accountable, and transparent AI systems, which involves addressing potential biases in AI algorithms or data, considering privacy issues, and ensuring AI is used for beneficial purposes. But these ethical considerations are imparted by humans, not innately adopted by the AI.

– Sapien, a GPT-4 powered AI assistant

Kieran Lindsay
FOUNDER, THE TRANSPARENCY PROJECT | TRANSPARENCY-PROJECT.AI

The Transparency Project's Submission on Supporting Responsible AI

About The Transparency Project

The Transparency Project (The Project) is an educational and research initiative started by Sydney-based researcher and entrepreneur Kieran Lindsay.

The Project's aim is to educate on the culture-shaping potential of artificial intelligence. This has started with scrutinising and elucidating the moral complexity within artificial intelligence. Through education and consulting services, the Project champions ethically responsible AI development, fostering a future where AI decision-making is not just technically proficient but also ethically conscious and universally beneficial. The Project aims to inform and guide, fostering a culture of transparency, accountability, and inclusivity in AI practices. By challenging established norms and cultivating open dialogue, The Project seeks to shape long-term ethical considerations in AI design and implementation, thereby redefining our relationship with technology for the common good and betterment of all.

The Project welcomes the opportunity to contribute to the consultation on Australia's approach to supporting AI in an ethical, safe, and responsible manner.[1]

This submission seeks to contribute to the following questions:

11. What initiatives or government action can increase public trust in AI deployment to encourage more people to use AI?

17. What elements should be in a risk-based approach for addressing potential AI risks? Do you support the elements presented in Attachment C?

20. Should a risk-based approach for responsible AI be a voluntary or self-regulation tool or be mandated through regulation? And should it apply to:

a. public or private organisations or both?

b. developers or deployers or both?

The Project recognises that AI regulation is challenging, continuously evolving and multifaceted. While our interests in AI regulation extend beyond this submission, we anticipate that the issues contained within will receive less attention. As such, our submission will focus on the cultural and, more specifically, the moral impacts of artificial intelligence and automated decision-makers.

Executive Summary

As more and more Australians embrace the use of AI, and as organisations and school systems start implementing its use, AI will have an ever-increasing potential to impact our cultural and moral thought systems. While Australia's democratic system is well-placed to allow for informed and evidence-based decisions that reflect community values, the potential impacts of ubiquitous AI on Australia's moral fabric will take time and targeted research to assess.

[1] The views expressed in this submission are those of The Project and Kieran Lindsay's role as founder of The Project. They are not the views of Kieran's other employers, who may be making separate submissions.

The Project's ultimate recommendation is that further consultation and research on the moral and cultural impacts of generative AI be undertaken to address the broad considerations in this submission. More concretely, we propose the following:

11. What initiatives or government action can increase public trust in AI deployment to encourage more people to use AI?

Australia should *consider* regulations that specifically address the moral worldviews of AI systems, especially of LLMs that power widely used applications and those deployed in morally-sensitive areas (such as early-childhood education and care).

It is envisioned that giving AI applications the moral endorsement from Government or a trusted regulatory body will reduce potential concerns about the values and principles of these systems.

At the same time, providing regulatory certainty as to the acceptable values displayed by AI systems will give the creators of these systems greater confidence to strike the right balance on ethical safeguarding (as opposed to an overly cautious approach). Such an outcome would allow for more robust and powerful systems and open up a more comprehensive range of use cases, increasing the users of AI technology in Australia.

17. What elements should be in a risk-based approach for addressing potential AI risks? Do you support the elements presented in Attachment C?

Whether a risk-based approach to regulating morality in AI systems is warranted needs further consideration. However, higher-risk applications of AI, considering the potential harms from a moral point of view, can easily be identified.

For example, applications of AI in education and morally sensitive use cases such as care settings present significantly higher risks. Australia's youth being raised on AI tools that follow and teach the moral worldview of overseas corporations presents the genuine threat that Australia loses any moral individualism, resulting in our population trending towards the "global mean" (as will other nations that use these technologies). This is not necessarily a bad thing (as discussed below), *but* if it is considered a negative outcome of AI, the risks of AI's use in shaping Australia's moral fabric in different settings should be considered.

20. Should a risk-based approach for responsible AI be a voluntary or self-regulation tool or be mandated through regulation? And should it apply to:

a. public or private organisations or both?

b. developers or deployers or both?

Generally, The Project believes that Australia should adopt a responsive and flexible co-regulatory approach to AI. The Project advocates that Australia should legislate regulatory powers allowing a new, independent regulator to register industry-led codes of practice covering different use cases and risk levels of AI technology. The regulatory body, at a minimum, should possess strong information-gathering and enforcement powers. For codes to be registered, they should require regular and in-depth reporting from AI companies to maximise transparency.

To minimise regulatory burden and maximise innovation, there should be no presumption that codes must be Australian-centric or created solely for Australia. The

code registration criteria should allow for international codes to be approved – as long as they meet the level of safeguards that Australia sees fit to require.

Suppose it is determined desirable that Australia regulates that AI systems embody a distinctly Australian moral fabric. In that case, this should form part of the ***mandatory*** requirements for ***deploying*** an AI system in Australia. Whether this would be for all AI use or just in public organisations (e.g., in schools and hospitals) or under different criteria (e.g., the average age of users and/or the number of users of a system) requires further consideration.

The burden of regulating the morality of AI systems will depend on how nuanced Australia wants locally-used AI systems to reflect the country's moral fabric. Australia's values already overlap significantly with America, which is currently leading the way in LLM development.

If Australia is to dictate such a requirement, it will have to develop a standard against which the moral values of AI systems can be tested. This should be a closed standard to minimise the chance it can be gamified. It would be appropriate that the requirement for testing AI against such a standard sits with a dedicated AI regulatory body.

As this submission is concerned with the impacts of AI on end users and the general public, the onus should sit with those deploying the technology to ensure the systems being deployed meet the moral standard. However, as a technical problem, the developers will be required to implement a technological solution to ensure that moral standards are achieved.

Lastly, it must be emphasised that this issue should not be skimmed over or tossed aside as too academic or philosophical. Moral development and young people's development, in general, is a well-studied field of research that has demonstrated how susceptible young minds are to their moral environments. Australia and countries around the globe are proposing to embed AI systems throughout society, from early childhood education to end-of-life care. This will undoubtedly have an impact on our society's moral fabric.

Whether or not AI systems can exercise moral judgement is irrelevant. As Sapien said:

> *AI can be designed to respect and adhere to ethical guidelines or rules set by its developers. … these ethical considerations are imparted by humans, not innately adopted by the AI.*[2]

The problem is that what is considered ethical by the non-elected developers in countries overseas may not be shared by us Australians.

Ultimately, Australia must decide whether our unique and diverse moral fabric is worth maintaining.

Morality in the Machines

Whether or not artificial intelligence can be considered as possessing moral agency is an academic pursuit beyond the scope of this submission. A pragmatic approach must be adopted. If an AI decision-maker is faced with moral challenges, how it overcomes these challenges requires careful consideration. And this is not even necessarily

[2] See quote on title page.

restricted to *AI*-powered automated decision-makers. As highlighted by Oliver Bendel, a simple household robotic vacuum may be faced with a dilemma as to whether or not to suck up and kill a ladybird (or a spider, and whether this makes any difference).[3]

Crucially, this moral decision-making cannot just be programmed around by defaulting to human oversight in more critical or advanced applications. Whether or not something has reached the threshold that it becomes morally contentious enough to warrant human oversight is itself a moral choice that the AI decision-maker must make.

The issue of moral machines has two dimensions when considering potential regulatory approaches. These two dimensions arise from the question: who is the appropriate moral vector in autonomous decision-making applications? The autonomous system? Or, the human creator?

The answer is a fuzzy both. It is easily within a developer's capabilities to furnish an autonomous vacuum cleaner with the ability to identify an insect with computer vision and decide whether to vacuum it. In this scenario, the developer is making the choice as to whether and what insects ought or ought not to be killed via vacuum.

Advanced AI, however, presents more difficulties. The promise of AI-powered autonomous decision-makers lies in their abilities to make decisions beyond the *if* and *if else* statements within traditional programs. These programs are designed to make decisions that the developers have not considered or could not comprehend during their programming.

The most straightforward illustration of this is Google Deepmind's AlphaGoZero, a self-taught artificial intelligence program designed to play the cognitively challenging game Go. AlphaGoZero taught itself to make decisions that allowed it to defeat the world's best human Go players – a feat far beyond the Go-playing skills of the system's creators.[4]

As such, it is both the developer and the system itself that needs to be considered in any regulatory approach. It would not be enough to require developers to uphold certain moral values in their programs. Or to only approve using autonomous decision-makers from countries or companies with shared values.

If Australia is going to take advantage of AI in an ethical, safe and responsible way, all aspects of an autonomous agent's design that allows it to learn and make decisions need to be scrutinised and elucidated. This will be increasingly important as AI-powered autonomous decision-makers become more advanced and are given greater freedom to learn and make more ethically contentious decisions (such as autonomous weapon systems).

Ultimately, regulation needs to ensure AI systems are built morally and transparently from the ground up – it may not be sufficient to rely on post-hoc human safeguards.

Loss of Moral Agency

Moral agency – the ability to make moral choices – has always been considered a distinctively human trait. We rarely morally condemn a vicious dog that has just

[3] Oliver Bendel, 'LADYBIRD: The Animal-Friendly Robot Vacuum Cleaner' in *The AAAI 2017 Spring Symposium on Artificial Intelligence for the Social Good Technical Report SS-17-01* (2017) 2.
[4] Misselhorn (n 2) 33.

mauled a baby – instead, we condemn the human owners who allowed such an atrocity to happen.

Modern-day philosophers and legal theorists such as Roger Brownsword and Mary Midgley have argued that it is crucial for the moral development of humans that they are allowed to make moral choices.[5] Aristotle held a similar point of view when he declared that moral virtue, vital for a good life, was the result of habit.[6]

Some have argued that by delegating moral decision-making to AI machines, humans are missing out on this crucial practice of making moral decisions.[7] Just like a bodybuilder increasing muscle mass through repetitive weightlifting, moral development is hindered without the ability to choose to act virtuously. This has been termed in the literature as "moral deskilling".[8]

This is a tricky conundrum. On the one hand, we want AI to be safe and automatically have 'desirable' moral viewpoints built in. However, without being presented with opportunities to make moral choices, this may be detrimental to our human ability to exercise moral judgement. This is especially so for the application of generative AI within classrooms. If children are not presented with contrarian moral viewpoints and get practice in making moral judgements, this may be detrimental to their long-term development as competent and individual moral agents.

Without any regulatory or societal indicators regarding the level of ethical safeguarding considered appropriate in generative AI applications, companies have become increasingly cautious about what morally contentious areas the chatbots will wade into. As witnessed in groups such as the ChatGPT Reddit community, it is an increasing gripe among ChatGPT's userbase to see messages along the lines of:

As an AI language model … [I can't respond to this request due to some ethical concern…]

This level of cautious safeguarding is a prudent and arguably appropriate approach by the companies to maintain their social licences while regulatory and broader community attitudes adjust to these issues. Yet, at the same time, in increasing such safeguards, LLM models are reducing the capacity for users to exercise their moral agency, even in situations where such a moral dilemma is unsettled due to cultural, religious, personal and other reasons.

Of course, there are other opportunities in life to flex moral muscles,[9] however, as AI becomes increasingly embedded in everyday life, its impacts in this area should not be discounted.

What level of ethical safeguarding generative AI and other AI machines have built-in should form another critical consideration of AI regulation. This must involve a balance of safeguarding against harm, such as biases, but allowing a degree of human moral thought in whether or not to accept or challenge AI content or decisions.

---

[5] Roger Brownsword, 'In the year 2061: from law to technological management' (2015) 7(1) *Law, Innovation and Technology* 1; Mary Midgley, *Wickedness: A Philosophical Essay* (Routledge, 1984) 3. See also Penny Crofts and Honni van Rijswijk, *Technology, New Trajectories in law* (Routledge, 2021) 14.
[6] Aristotle, *The Nicomachean Ethics*, tr David Ross (Oxford University Press, 2009).
[7] Shannon Vallor, 'Moral Deskilling and Upskilling in a New Machine Age: Reflections on the Ambiguous Future of Character' (2014) 28(1) *Philosophy & Technology* 107.
[8] Ibid.
[9] Paul Formosa and Malcolm Ryan, 'Making Moral Machines: Why We Need Artificial Moral Agents' (2020) 36(3) *AI & SOCIETY* 839.

Moral Homogenisation

Hand-in-hand with the issue of moral deskilling is the idea of moral homogenisation. Moral homogenisation refers to the potential for ubiquitous AI use to result in a loss of diversity in moral worldviews.

If we are building machines that must make moral decisions, it will be necessary that these machines express a moral stance, whereas, in humans, such dilemmas have been left up to individual choice. The classic example is the application of the trolley problem (and various similar dilemmas) to autonomous vehicles. If an autonomous car is faced with a choice to either run over a mother and a child or swerve and kill the elderly passenger, which will it choose? Whether or not a human hardcodes the 'correct' decision for each dilemma into the program or the system is left to learn and decide the correct outcome (e.g. like AlphaGoZero), the result will be a system that will make the same choices time after time.[10] Different autonomous decision-makers may have the capabilities to make different choices, but at least for commercial products, each decision-maker will be designed to make choices that are most appealing to the largest group of people (to maximise commercial opportunities).

It also needs to be considered who is in control of making these decisions. Currently, the moral worldview of AI systems is controlled by the companies creating the technology. For generative AI applications, the developers of the LLMs are setting the moral worldviews, ethical safeguards and decision-making processes of such systems. These companies are likely driven by two influences, the general moral worldview of the community they are part of and, probably more influentially, the commercial appeal of a worldview that appeals to the largest number of revenue-driving users.

This may or may not be problematic. Reducing moral contention by exposing users to a more coherent and unified moral worldview may result in less friction between people and groups. This could be seen as the next step of globalisation, where the differences between states become less defined as global populations are raised on ethical machines developed by a small subset of companies that are expressing a similar worldview. Friction between nations reduces, and the world becomes more prosperous as countries collaborate more closely.

On the other hand, such an approach will see a loss of nations' cultural and moral identity. The problem arises not around ethical dilemmas that are highly universal but instead in cases where the ethics could be considered more relative. Most Australians seem happy to let go of a small amount of freedom to ensure guns are tightly controlled. Should Australians, especially Australian children with less developed critical thinking skills, be surrounded by AI models built and trained by American companies that do not necessarily share this view?

It is a tricky position, and whether Australia imposes a requirement for a distinct, Australian moral worldview on the LLMs and generative AI applications in the country should be a serious consideration. Only with the development of LLMs trained in Australia, pursuant to Australian ideals, or the regulation of the moral worldviews of international products can Australia maintain control over the ideals and values expressed in AI applications in the country. This is not a far-fetched idea either. OpenAI itself recently called for grant applications from those looking into the

[10] There have been suggestions that in such ethical dilemmas, the autonomous agent should select an outcome at random. See, Liang Zhao and Wenlong Li, '"Choose for No Choose"—Random-Selecting Option for the Trolley Problem in Autonomous Driving' (Conference Paper, Proceedings of the 9th International Conference on Logistics, Informatics and Service Sciences, 10 July 2020).

'democratisation' of AI, including the question: '[s]should AI by default reflect the persona of a median individual in the world, the user's country, the user's demographic, or something entirely different?'[11]

LLMs trained by American companies power the most popular and widely used generative AI tools. While Australia shares a similar worldview to America, this is not going to be universal, and as more solutions from different countries become available and their use in schools and other settings increases, this issue will only grow. China, so far, is the clearest example of dealing with this issue in that it has put forth regulatory guidelines for LLM development, including that content generated must align with the country's socialist values.[12]

Regulating the moral worldview of AI systems may also result in Australians trusting and adopting AI systems more readily. If Australians know and are confident the system is approved to meet the values of the Australian community, another potential barrier to adoption is removed. This is even more relevant for use in situations such as education or care settings where scepticism of the values these systems espouse may be preventing uptake and preventing the benefits of this technology from being realised.

Giving regulatory certainty over morality in AI systems may also give Australians access to more powerful AI systems. If the creators are given clarity and a way to measure that their systems comply with Australia's values, they can remove overly cautious safeguards that may be hindering the power and application of these systems.

Does Moral by Design Equate to Safe AI?

Ensuring AI systems are moral by design has been put forth as one approach to mitigate potential harm from AI. The idea is that an AI with a strong moral foundation is less likely to cause harm. This is akin to how we often perceive individuals with strong moral virtues as less likely to engage in harmful or criminal behaviour. While this may be the case to an extent, such an idealist approach to morality is not necessarily congruent with 'safety' in AI. What Australia wants regarding safety may not always align with what we consider the morally appropriate choice.

For example, consider the context of elder care. If an AI system is programmed solely with a 'safety first' principle, it might limit the activities of an elderly person to avoid any potential health risks. For example, should an AI system prevent an older adult in its care from an evening glass of wine or a second serving of pizza?[13] How about preventing a terminal patient from enjoying a cigarette?

Defining what is meant by 'safe and responsible AI' and how this balances with what is considered right and wrong will become necessary if AI is to be deployed in such situations.

However, AI systems with a solid moral constitution may provide far superior and safer decision-makers in situations where human emotions and biases may be harmful. While bias in AI is a real and serious issue, these biases remain consistent and can,

[11] Wojciech Zaremba et al, OpenAI 'Democratic inputs to AI' (Announcement, 25 May 2023) <https://openai.com/blog/democratic-inputs-to-ai>.
[12] Josh Ye, 'China proposes measures to manage generative AI services', Reuters (online, 11 April 2023) <https://www.reuters.com/technology/china-releases-draft-measures-managing-generative-artificial-intelligence-2023-04-11/>
[13] Misselhorn (n 2) 44.

therefore, be more easily managed. This is far less achievable in managing human biases across different individuals. For example, in situations of the application of deadly force, a human with unknown biases and prejudices, such as past trauma or racist disposition, is a far more unpredictable decision-maker than an AI system. Further, an AI system with a strong moral sense may also be less susceptible to biases arising from its training data. Combining this with the ability for AI to make decisions in a fraction of the time it would take humans, moral AI systems may be considered safer and superior for certain applications.

The Difficulties of Moral AI Systems

Regulating the morality of AI systems is a complex endeavour, presenting challenges not just from a conceptual standpoint but also concerning the practical implications of imbuing machines with moral perspectives. These challenges are multi-faceted and interwoven with societal and technological concerns.

The Challenge of Consensus on Moral Frameworks

Firstly, there is a significant challenge in achieving consensus on which moral principles to express in AI systems. As with any multicultural and diverse society, Australia's populace holds a myriad of moral perspectives informed by cultural, religious, socio-economic, and personal factors. Distilling these diverse views into a universally acceptable moral framework for AI is a daunting task.

However, this is not a barrier. Australia has been able to regulate and legislate for morally contentious issues, such as the passing of Maeve's law last year to allow for mitochondrial donation.[14]

Upholding strong democratic ideals to ensure that the moral principles adopted by AI systems reflect the views of the broader community will be essential. Public consultation should be inclusive and representative, involving individuals from different backgrounds, cultures, age groups, and socio-economic statuses. Regular reviews and updates to this moral framework can also help keep it in line with evolving societal norms and expectations.

The Challenge of Measuring AI Morality

The second challenge lies in quantifying and measuring AI system morality. Morality is abstract and subjective, unlike more tangible aspects of AI performance, such as processing speed or accuracy. Therefore, determining how 'moral' an AI system is becomes problematic.

One potential approach is to develop a set of quantifiable metrics that reflect the AI system's adherence to the moral principles defined through public consultation. The Project is currently testing the feasibility of such an approach by measuring different aspects of morality through the API endpoints of publicly available LLMs.

However, these metrics and audits should not replace the need for ongoing scrutiny and qualitative assessments. User feedback, case studies, and public debates can provide valuable insights into the real-world ethical performance of AI systems, capturing nuances that quantitative metrics might miss.

[14] *Mitochondrial Donation Law Reform (Maeve's Law) Act* 2022 (Cth).

Incorporating AI Morality in Practice

Translating moral principles into practical, machine-readable guidelines is another significant challenge.

Machine learning techniques, such as reinforcement learning, can be used to 'train' AI systems to make morally appropriate decisions. However, this training must be carefully managed to prevent the system from learning undesirable behaviours. Transparency in training AI systems to meet moral standards, plus rigorous testing and scrutiny, is a minimum.

Conclusion

Morality in AI should not go unconsidered in Australia's approach to responsible and safe AI regulation. However, regulating the morality of AI systems in Australia (or even considering whether we should) would undoubtedly be a complex and challenging process that goes beyond the scope of setting technical guidelines or practical principles and standards. Yet the potential for ubiquitous AI to shape the moral culture of Australia demands that this is a valuable exercise to undertake and one which has the potential to accelerate the realisation of the benefits of AI in Australia.

As such, The Project recommends that the issue is given further consideration. With further inclusive public consultation, careful definition, development of the ability to reliably measure AI morality and the practical implementation of moral guidelines, it is feasible to create AI systems that reflect and uphold Australia's shared moral values (which may include the desire to have no moral values in AI at all!). This process will make AI systems safer and ensure that they serve the diverse needs and values of Australian society.