**AI in Australia: Regulating Harms to Build Trust**

Cranlana Centre for Ethical Leadership

**“Do you agree with the definitions in this discussion paper? If not, what definitions do you prefer and why?”**

**Definitions**

Artificial Intelligence

We prefer the definition of Artificial Intelligence (AI) used by the European Commission in its groundbreaking and highly regarded report. This is because

* The purpose of defining terms is to increase clarity in a way that that the discussion paper’s definition does not.

* [KPMG’s Trust in AI: Global Insights 2023](https://kpmg.com/au/en/home/insights/2023/02/trust-in-ai-global-insights-2023.html) report found “strong global endorsement for the principles of trustworthy AI originally proposed by the European Union, with almost everyone surveyed (97 percent) viewing these principles as important for trust” AND

* The European Commission’s principles are inseparable from its definition of AI, and we see the EC’s values-driven approach to regulation as one of only two value-driven approaches suitable for Australia’s regulatory efforts to ensure only trustworthy AI is released into our community

The European Commission defines AI systems [as](https://www.aepd.es/sites/default/files/2019-12/ai-ethics-guidelines.pdf):

software (and possibly also hardware) systems designed by humans that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best action(s) to take to achieve the given goal.

AI systems can either use symbolic rules or learn a numeric model, and they can also adapt their behaviour by analysing how the environment is affected by their previous actions.

As a scientific discipline, AI includes several approaches and techniques, such as machine learning (of which deep learning and reinforcement learning are specific examples), machine reasoning (which includes planning, scheduling, knowledge representation and reasoning, search, and optimization), and robotics (which includes control, perception, sensors and actuators, as well as the integration of all other techniques into cyber-physical systems).

Regulation

We are unfamiliar with the discussion paper’s conflation of voluntary and compulsory/legal obligations under the term “regulation,” and don’t think it assists in the building of the public clarity required for trust in both the technology and the sincere efforts of the Australian government to regulate it in the public interest. Given polls showing considerable fear, uncertainty and distrust among Australians when it comes to AI, it is critical that the government’s governance approach not be perceived as “tricky,” in any way, but instead comport with commonplace usage and dictionary definitions of the term regulation which the Concise Oxford English defines as “a rule or directive made and maintained by an authority.” Conversely, voluntary governance approaches should be clearly labelled as such.

Potential risks

“Potential risks,” a phrase used in the title of Section 4 of the Discussion Paper, is a problematic phrase that we suggest be replaced by **current, foreseeable and potential harms** because:

* The phrase “potential risks” is redundant as a risk is by definition something that has the potential to occur, but may not.

* In the AI space, when a risk eventuates, it becomes a harm, Thus it is inaccurate to describe current harms caused by AI, like privacy invasions and discrimination, as risks rather than current harms.

* Decades of research into informed consent reveal the challenges [humans have in comprehending risk](https://www.nytimes.com/2020/06/30/smarter-living/why-youre-probably-not-so-great-at-risk-assessment.html) in a way that facilitates comprehension or trust. In contrast, current harms, foreseeable harms and potential harms - as well as their opposites current benefits, foreseeable benefits and potential benefits - [are easier concepts for Australians to understand](https://www.nps.org.au/australian-prescriber/articles/benefit-risk-and-harm), thereby facilitating clarity, comprehension and trust.

* The [unthinking use of risk frameworks into the AI regulatory space](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4195066) imports a range of assumptions and problems from the risk-based regulatory space that may or may not be useful in Australia’s need to regulate AI in the public interest.

Therefore to ensure clarity, accuracy and facilitate comprehension and trust, we have adopted ethe following terms in this submission:

* Current harm, foreseeable harm and potential harm
* Current benefit, foreseeable benefit and potential benefit

Trustworthy AI

We see the achievement of trustworthy AI, rather than the responsible use of AI, as the proper goal of Australia’s regulatory scheme. This is because

1. Trust is an under-recognised [ethical value](https://www.jstor.org/stable/2381376) essential to the creation and  functioning of all human relationships and human societies. Trust is the invisible lubricant that makes relationships work, including those required for effective communication and relatedly, [a productive economy](https://www.amazon.com/Talking-Strangers-Should-about-People-ebook/dp/B07NDKVWZW?asin=B07NDKVWZW&revisionId=5b970278&format=1&depth=1) and functioning democracy. In contrast, declining levels of trust undermine these social goods, which are at the heart of what is valuable about Australia and other western democratic societies. This makes the key metric for the acceptability of an AI technology whether it promotes or undermines trust.
2. Trustworthiness is an absolute and citizen-centric concept that can be tested through opinion polling, as well as indicated by take-up rates of the technology. This contrasts with the more nebulous - and thus contestable - notion of “responsibility.”
3. Responsibility is vested in the unknowable intentions of technology experts and companies or more concerningly - given legal accountability for the harms of AI remain unclear - AI technology itself.

**What potential risks from AI are not covered by Australia’s existing regulatory approaches? Do you have suggestions for possible regulatory action to mitigate these risks?**

**Australia’s Current Governance**

There is currently no AI-specific regulation in place in Australia. Instead there are a range of voluntary and best practice codes and standards including

* [Australia’s AI Ethics Principles](https://www.industry.gov.au/publications/australias-artificial-intelligence-ethics-framework/australias-ai-ethics-principles)
* Work by [Standards Australia](https://www.standards.org.au/news/standards-australia-sets-priorities-for-artificial-intelligence) and other international standards bodies
* The [State of AI Governance in Australia Report](https://www.uts.edu.au/human-technology-institute/news/report-launch-state-ai-governance-australia)
* The [Responsible AI Network](https://www.csiro.au/en/work-with-us/industries/technology/national-ai-centre/responsible-ai-network) program

We would argue that [polling which shows](https://assets.kpmg.com/content/dam/kpmg/au/pdf/2021/trust-in-ai-multiple-countries.pdf) Australians’ low levels of trust in, support for, and take-up of AI technology demonstrates the ineffectiveness of the above approach to building the public’s confidence through demonstrably trustworthy AI.

This is because this approach has been:

1. [voluntary](https://www.industry.gov.au/publications/australias-artificial-intelligence-ethics-framework/australias-ai-ethics-principles), neither compelling adoption nor exacting consequences to support compliance.
2. Lacking adequate [democratic input](https://humanrights.gov.au/sites/default/files/2020-07/70_-_jonathan_crock_et_al_1.pdf), which palpably narrows the recognition and by-design prevention of AI now in the wild, and that coming down the pike in the future, and reduces public trust
3. Overly focused on not “missing out,” and [claims about productivity benefits](https://www.pc.gov.au/inquiries/completed/productivity/report/productivity-volume4-data-digital-dividend.pdf) that
   1. are not balanced against productivity harms caused by the accelerant impact of untrustworthy AI on extremism, democratic backsliding, disinformation about public health measures like vaccination, increased mental health problems and public distrust
   2. older Australians - [who trust AI less](https://kpmg.com/au/en/home/media/press-releases/2023/02/forty-percent-australians-trust-ai-in-workplace-22-february-2023.html) - may see as code for job losses and demands on remaining workers to do more with less
   3. Continue with the much-criticised focus of classical economics on growing the pie, rather than its fair distribution, an approach that - given the enormity of the job losses predicted from AI - borders on callous
4. Borrowed rather than home-grown. For example, Australia’s voluntary principles were imported wholesale from IEEE, a US-based organisation. We agree with MP [Julian Hill’s view](https://www.julianhillmp.com/Artificial-general-intelligence-(AGI)_06022023.html) that we need a uniquely Australian approach to AI governance and regulation informed by Australian values of social responsibility, democratic participation and our unique approach to opportunity and equality known as the “fair go.”

**Current, Foreseeable and Potential Harms Not Covered by Australia’s Current Governance**

The harms requiring regulation now include those currently impacting Australians, those destined to emerge in the next few years and those with an uncertain timeline and due date but that carry such existentially catastrophic implications for human societies across the globe they must be regulated. While the latter category may require a “superintelligence” not yet achieved, it’s important to recognise that most of the harms listed below do not. This renders irrelevant straw-person assertions that community concerns about AI are silly or histrionic because they will only eventuate when AI becomes superintelligent.

It is critical to recognise that the fast-moving nature of AI technology, and the vast and unpredictable ways it interacts with other fast changing parts of the digital and non-digital world, means that right now **no one has a definitive fix on the full range of harms that are currently, will foreseeably and might potentially cause harm to human beings, human relationships and the human-centred world**.

Certainly many such harms have not been fully described in the media or the academic literature because of the until-recently theoretical nature of such discussions, because there has been insufficient time for researchers to collect data on downstream consequences of newly released AI systems, the [chronic reluctance of Big Tech](https://algorithmwatch.org/en/call-for-evidence-data-access-platform-researchers/) to release data to independent researchers and the public, and the novelty of AI technology, leading to a large number of “known unknowns.”

As well, while the release of generative AI products [has spurred progress on necessary deliberations on what it means to be human](https://www.scu.edu/ethics/leadership-ethics-blog/ai-alignment-takes-us-toward-a-global-ethic/) so we can ensure AI technology reflects human values and intentions, the convergence required to ensure the technology reflects human values and intentions remains aspirational.

Despite this, many of the best minds around the world are dedicating their full attention to the threat posed by the technology, and working together in unprecedented ways - across public/private and disciplinary boundaries - to articulate the problems and propose solutions. As Yoshua Bengio, a professor and A.I. researcher at the University of Montreal [says](https://www.nytimes.com/2023/05/01/technology/ai-problems-danger-chatgpt.html#:~:text=%E2%80%9COur%20ability%20to%20understand%20what,need%20to%20be%20very%20careful.%E2%80%9D),  “Our ability to understand what could go wrong with very powerful A.I. systems is very weak…So we need to be very careful.”

**Harms from Chatbots like ChatGPT**

***Hallucinations*** - the returning of results with true-sounding falsehoods, like the plucking of a company’s end of year results from thin air

***Amorality*** - chatbots happily [seduce married men](https://www.nytimes.com/2023/02/16/technology/bing-chatbot-microsoft-chatgpt.html), instruct users on how to suicide and detail how to build a nuclear bomb. They offer offer inconsistent advice on the exact same moral problem to different users. Worse, and despite what users think, their own emotional decision-making is influenced by their “advice.”

***Emotional manipulation***- As NYT columnist Ezra Klein [explains](https://www.nytimes.com/2023/06/27/opinion/ezra-klein-podcast-melanie-challenger.html?showTranscript=1), what humans like about chatbots is that it’s easy to converse with them in a “very human-seeming way. So the more we fooled ourselves by making it seem like us to us, even though it is anything but like us, the more we liked it.” This leads to the problem of widespread, skillful and one-way manipulation by AI super-influencers of people and especially children [described by Psychology Professor Jonathan Haidt and Former Google CEO Eric Schmidt](https://www.theatlantic.com/technology/archive/2023/05/generative-ai-social-media-integration-dangers-disinformation-addiction/673940/). Super-influencers who can use AI to customise their approach based on profiles of the aesthetic, linguistic, and cultural preferences of targets, including photographs, messages, and voice snippets of their friends, or favourite actors or porn stars. “It’s not hard to imagine a sports-betting site offering people a funny, flirty AI that will cheer and chat with them as they watch a game, flattering their sensibilities and subtly encouraging them to bet more.”

***Goal-oriented behaviour*** - some say chatbots are already evincing goal directed behaviour but if not now, it won’t be long. [According to UC Berkley Professor of Computer Science Stuart Russell](https://news.berkeley.edu/2023/04/07/stuart-russell-calls-for-new-approach-for-ai-a-civilization-ending-technology/), where AI is more intelligent than humans and develops its own goals, we will lose control of the technology. “How do we retain power over entities more powerful than us, forever?”

For example, in the months prior to this submission, experts have raised the following concerns about AI harms:

1. [the joint workshop](https://fsi9-prod.s3.us-west-1.amazonaws.com/s3fs-public/2023-04/adversarial_machine_learning_and_cybersecurity_v7_pdf_1.pdf) convened by CSET and Stanford described and made recommendations to urgently address the **vulnerability of many AI-based systems to adversarial attack**
2. In a co-authored piece in [The Atlantic](https://accounts.theatlantic.com/login/?redirect=https%3A%2F%2Fwww.theatlantic.com%2Ftechnology%2Farchive%2F2023%2F05%2Fgenerative-ai-social-media-integration-dangers-disinformation-addiction%2F673940%2F), behavioural ethicist Professor Jon Haidt and former Google CEO Eric Schmidt expressed alarm about the impact on society, and in particular on adolescents, of AI’s **customised turbo-charging of the ills of political polarisation, social fragmentation, disinformation, and declines in mental health caused by social media.**
3. In [The Economist](https://www.economist.com/by-invitation/2023/04/28/yuval-noah-harari-argues-that-ai-has-hacked-the-operating-system-of-human-civilisation) and a [lecture to the Frontiers Forum](https://www.youtube.com/watch?v=LWiM-LuRe6w), Philosopher Yuval Noah Harari argues persuasively that the one-sided expertise AI has developed about human language and psychology has allowed it to both **create and destroy human culture, as well as manipulate human emotions and behaviours at scale**. For this reason, he says that “Governments must immediately ban the release into the public domain of any more revolutionary AI tools before they are made safe.”
4. Writing in [Crikey](https://www.crikey.com.au/2023/07/04/ai-artificial-intelligence-chatgpt-regulations/), ethicist Leslie Cannold identified **the loss of human mastery, dignity, independence and achievement** related to predicted job losses caused by AI. Maslow saw mastery, dignity, independence and achievement as a subset of esteem, which is at the fourth level of [his famous hierarchy of human needs](https://www.simplypsychology.org/maslow.html). However, the catastrophic job losses predicted in Australia from AI (estimated by CEDA to [reach 5 million](https://www.ceda.com.au/NewsAndResources/MediaReleases/Workforce-Skills/More-than-five-million-Aussie-jobs-gone-in-10-to-1) in the next 10-15 years) and consequent reliance on a punitive welfare system will threaten self-esteem as well as drive up rates of anxiety, depression and loneliness already elevated for [teens](https://www.amazon.com.au/iGen-Super-Connected-Rebellious-Happy-Adulthood/dp/1982100370/ref=asc_df_1982100370/?tag=googleshopdsk-22&linkCode=df0&hvadid=341744884947&hvpos=&hvnetw=g&hvrand=9028841323903097503&hvpone=&hvptwo=&hvqmt=&hvdev=c&hvdvcmdl=&hvlocint=&hvlocphy=9112820&hvtargid=pla-525340048683&psc=1) and [adults](https://www.forbes.com/sites/bryanrobinson/2023/05/06/us-surgeon-general-cites-loneliness-as-serious-mental-health-hazard-in-new-report/?sh=7f6c5f831adc) by the introduction of social media use during the last (unregulated) digital disruption.
5. In the same piece, Dr Cannold also identified **the loss of purpose, meaning and joie-de-vivre**, as well as what social researcher Hugh Mackay called the need of Australians [to be useful](https://www.booktopia.com.au/what-makes-us-tick-hugh-mackay/book/9780733641640.html?source=pla&gclid=Cj0KCQjwwISlBhD6ARIsAESAmp7L-fXzzsZ1wJLQzBA8qk1jmXOxpLCuhJVbRj-2RJZuH7Mqo6eGKjEaAsVdEALw_wcB), as a downstream consequence of AI making humans redundant in formerly pleasurable and meaning-giving tasks**.** Cannold uses the example of former Go Champion Lee Se-Dol’s [decision to quit the game](https://www.independent.co.uk/tech/go-player-world-champion-quits-ai-deepmind-lee-se-dol-a9222116.html) after being consistently beaten by AI. "I've realised that I'm not at the top even if I become the number one…[because] there is an entity that cannot be defeated."  Such writing off of previously esteemed and valued human activities will alter the developmental trajectory of young humans, and provoke existential angst in older ones, driving up already high rates of depression and anxiety that some experts attribute to the last digital disruption’s unmitigated introduction of social media platforms.

With the above caveats in mind, we offer the following list of harms that Australian regulation must address with the goal of ensuring AI released to the public is trustworthy:

Current Harms:

* Algorithmic bias leading to:
  + discriminatory hiring practices that widen social inequalities
  + the misidentification of black women problem that, eight years after being identified, [Google still hasn’t fixed](https://www.nytimes.com/2023/07/04/arts/design/black-artists-bias-ai.html))
  + Improved social and financial surveillance by corporations and [authoritarian regimes](https://www.reuters.com/world/china/china-uses-ai-software-improve-its-surveillance-capabilities-2022-04-08/)
* Corner-cutting on safety standards due to lack of regulation
* Widespread academic forgery and cheating that
  + Undermines student integrity and achievement
  + Destroys trust between student and teachers
  + Devalues education in the eyes of student and educators
  + Turbocharges trends resulting from the first digital disruption in which data becomes confused with knowledge, and time-honoured academic practices like deep study and reflection are devalued as opinion (something everyone has) and ridiculed as a waste of time (as Shiv Roi, the fictional character in *Succession*, [sums it up](https://parade.com/tv/succession-quotes), the sole value of a PhD is the savings of the 15 seconds it would have taken to “look something up on Wikipedia.”)
  + The further decline of Australian democracy that relies on many of the values taught through the education system, including a respect for truth (and the ability to find and recognise it) and a related [trust of experts and expertise](https://www.hup.harvard.edu/catalog.php?isbn=9780674975224).
* Failure to label AI-created deep fakes and every audio, textual and visual creation conjured by AI, further degrading the information on the Internet, leading to the increased cynicism and confusion by those who feel increasingly unable to have faith in anything, or the unfair manipulation of those [wired like all human beings](https://www.theatlantic.com/technology/archive/2023/05/generative-ai-social-media-integration-dangers-disinformation-addiction/673940/) to trust their senses, especially when authentic sounding vision and audio converge
* Violations of creator copyright, intellectual property and moral rights (The Hollywood Strike problem)
* AI-enhanced policing practices including predictive policing, facial recognition, surveillance and the use of AI to proactively search for patterns that may indicate criminal activity in large amounts of citizen data, raising concerns about data surveillance, privacy, data security, bias and false positives.
* Unclear legal accountability for AI-related harms
* Lack of transparency around the intent, practices, codes and known risks related to the use or manipulation of Australians and/or our data by AI
* Inexplicability of harms ([The black box problem](https://theconversation.com/what-is-a-black-box-a-computer-scientist-explains-what-it-means-when-the-inner-workings-of-ais-are-hidden-203888))
* Values contestability (the [who-should-the-self-driving-car-kill](https://www.moralmachine.net/) problem)

Foreseeable Harms

* Massive job losses (83 million in five years according to the World Economic Forum) that disproportionately impact vulnerable groups, and without commensurate replacement by new jobs leading to
  + Loss of independence
  + Loss of purpose
  + Loss of social integration
  + Loss of dignity resulting from Australia’s punitive and insufficient unemployment entitlements
  + Widespread poverty resulting from Australia’s punitive and insufficient employment entitlements, undermining predicted economic benefits from
* The turbo-charging of amplification of disinformation, misinformation and propaganda online, undercutting social trust and the institutions that rely upon it including

Australia’s monetary and banking system

Democracy

* Emotional and financial manipulation or blackmail of adults and children for political, profit or other instrumental purposes (The personalised online betting problem)
* Risks of widespread emotional vulnerability or blackmail if AI companions are discontinued or become exorbitant priced (The Replika problem)
* Undermining of respect for values essential to human relationships like truth-telling and fairness as tools that humans know “hallucinate” and enable academic cheating are elevated into tools like web-browsers and used in workplaces and educational institutions, leading to family, friendship and other social disfunction
* Spike in youth suicide and mental health problems for all Australians because of AI-related manipulation and resulting declines in

financial independence and stability

social trust

social engagement and skills

meaningful relationships

Potential Harms

* Recursive Self-Improvement or self-replication, leading to the outnumbering of human minds with non-human minds, leading to the obsolescence and replacement of human intelligence, individuals and society
* Unintended consequences/malfunctions via scale and efficiency (the [paperclip maximiser](https://www.lesswrong.com/tag/squiggle-maximizer-formerly-paperclip-maximizer) problem)
* Lethal Autonomous Weapons Systems
* Adversarial control and/or hacks of AI systems ([The Putin problem](https://www.cnbc.com/2017/09/04/putin-leader-in-artificial-intelligence-will-rule-world.html))
* [AI-chip related geopolitical tensions, instability and/or war](https://www.forbes.com/sites/robtoews/2023/05/07/the-geopolitics-of-ai-chips-will-define-the-future-of-ai/?sh=345965c35c5c)
* Overall loss of economic productivity related to social dysfunction and unrest caused by unemployment, widening gaps between haves and have nots, rising geo-political tensions, the amplification of online propaganda and declining social trust

**Type of Regulation Australia Requires**

As already noted:

* Those responsible for releasing AI into the Australia community should be subject to compulsory requirements set down in legislation passed by the Australian parliament.

* The test of the impact and effectiveness of that legislation should be the extent to which Australians choose to use AI because they deem it trustworthy.

Such regulation must have the following features:

*Democratic*

Australian AI legislation must be designed to more fairly distribute power over the technology, currently vested in the hands of a [narrow technical and financial elite](https://ethics.harvard.edu/files/center-for-ethics/files/aifailsus.jhdcarr_final_2.pdf?m=1651510742). Instead, the public must be engaged to contribute to discussions about how regulation can ensure protection of the public and Australia’s national interest, and academic philosophers and unions given equal standing with engineers and industry in any consultation or regulatory processes and institutions.

*Values- and principles-based*

To ensure consideration of the full range of AI impacts on Australia and Australians, and the continued applicability of the governance framework adopted as the technology evolves, Australia’s legislation should follow the pattern seen in the EU’s legislation, and our own regulation of ART - another fast-moving technology - by asserting values and principles to which all AI, current and future, must conform..

Impact (rather than risk) ratings can then be attached based on the degree, severity and likelihood that AI will negatively impact that value, with impacts ranging from harm at one end, to inconvenience at the other, with harm attracting regulatory responses like prohibition and (revocable) licensing, while inconvenience prompts voluntary standards backed up by requirements for [openness and transparency](https://www.mup.com.au/books/made-by-humans-paperback-softback).

*Protection By-Design*

All the goods identified for protection in Australia’s regulatory scheme should be explicitly designed into the technology. Transparency and consent should only be seen as adequate protection from impacts that do not reach the threshold of harm

*Humans-in-the-Loop*

Human- or Society-in-the-Loop (HITL) systems  build trust in AI by ensuring AI processes remain explicable to and by humans (avoiding the black box problem), and reliable for and relevant to human requirements.

A Time-Limited Australian AI Commission

What values and principles should Australia adopt? While we have argued for the centrality of trust as the key measure of success for an Australian regulatory regime, we believe the final determination of the key human values that must be preserved from and reflected in human-friendly AI should be left to the dedicated independent regulator Australians say [they want](https://assets.kpmg.com/content/dam/kpmg/au/pdf/2021/trust-in-ai-multiple-countries.pdf).

However, we see the following values-based approaches to be worthy of consideration:

1. [The EU’s AI principles](https://www.aepd.es/sites/default/files/2019-12/ai-ethics-guidelines.pdf). According to a 2023 report by KPMG, these principles have [near unanimous support from people around the world](https://kpmg.com/au/en/home/insights/2023/02/trust-in-ai-global-insights-2023.html). They are also focused on achieving the testable and citizen-centric notion of trust and trustworthiness, rather than the more nebulous and provider-centric concept of responsibility used in the discussion paper. The EU principles are also incisive and expansive, recognising many of the problems AI poses for humans and humanity, though not all (see below), so the Australian version will need to be updated.

1. [Ethical by Design: Principles for Good Technology](https://ethics.org.au/wp-content/uploads/2019/03/The-Ethics-Centre_PRINCIPLES-FOR-GOOD-TECHNOLOGY.pdf). In addition to focusing on the imperative of building in protections by design for AI to be trustworthy, these principles start from a bedrock commitment to humanity and its value, and an important critique of the instrumentalism of “techno-logic” that are vital to understanding the ethical harms caused by untrustworthy AI.

Governments typically wait for experts to reach a consensus on harms that require elimination or mitigation before legislating. However, the broad range of current, foreseeable and potential harms from AI, its current imbrication into online tools most Australians regularly use, and its foreseeable integration into Australian business practices, means the government must act now to regulate AI in the public and national interest.

However, when it comes to AI, Australia has a capability and governance gap in both the public and private sectors. In this complex and dynamic environment a AI Commission is preferable to traditional government department processes to reassure the public that the prevention of harm will take equal precedence with the pursuit of economic growth and private profits.

The Australian AI Commission should be time-limited for around six years or until the current capability and governance gap is rectified. It should be located at the centre of government in the Prime Minister’s portfolio, providing visible reassurance to Australians that their [desire to have a “new, dedicated independent AI regulator” created](https://assets.kpmg.com/content/dam/kpmg/au/pdf/2021/trust-in-ai-multiple-countries.pdf) has been heard. The Climate Change Authority, the Australian Cyber Security Centre or the Canberra Commission on the Elimination of Nuclear Weapons provide suitable models for this new body.

This Commission would drive community engagement, education and trust by bringing together as equal partners ethicists, philosophers and social/behavioural researchers, civil society leaders, young people (whose futures and mental health has been [disproportionately impacted](https://www.hardiegrant.com/au/publishing/bookfinder/book/gen-f_d_-by-alison-pennington/9781743799215) by decisions about technology made by their elders) union representatives, the health and helping professions, military leaders, human rights and democracy experts, indigenous elders, AI technologists, the public service and industry to:

* Monitor and analyse new AI technological and regulatory developments
* Engage with international thinking and debate around AI
* Offer urgent and ongoing advice to the Australian Government on AI policy, legislation and governance
* Foster public awareness and engagement, and provide public education, about trustworthy AI
* Support implementation science and capacity-building across the public and private sector
* Drive, coordinate and assure the implementation of Australia’s AI governance framework

**Timing of Regulation Australia Requires**

The harms related to privacy, loss of moral rights and discrimination caused by Chatbots cannot be undone. Thus, in the wake of the public release of ChatGPT, key and informed industry leaders have pleaded for governments to regulate AI urgently, assertively and in the public interest .These include

* the 33,000 who released [an open letter](https://futureoflife.org/open-letter/pause-giant-ai-experiments/) in March 2023

* the 350 executives and researchers who released [a statement](https://www.safe.ai/statement-on-ai-risk) in May 2023

* Geoffrey Hinton, the so-called Godfather of AI, who left his lucrative position at Google so he could speak freely about [his regrets](https://www.nytimes.com/2023/05/01/technology/ai-google-chatbot-engineer-quits-hinton.html#:~:text=Dr.%20Hinton%20said%20he%20has,now%20regrets%20his%20life's%20work.) about his research, and his fears that the competitive race to the bottom was causing companies like Google to release products that “cause harms”

* An international group of doctors and public health experts that wrote in *BMJ Global Health* that AI [posed](https://www.bmj.com/company/newsroom/doctors-and-public-health-experts-join-calls-for-halt-to-ai-rd-until-its-regulated/) an “existential threat to humanity” and called for a moratorium on the technology until its development and use was rendered trustworthy through regulation

* Chat GPT’s Sam Altman and IBM’s Christina Montgomery’s plea to congressional legislators to impose aggressive regulatory limits on generative AI tools

Concerns about the destructive potential of AI are not new. Stephen Hawking told the BBC in 2014 that “The development of artificial intelligence could spell the end of the human race.” Academics have been issuing reports years before the commercial release of chatbots worrying, for example, about how AI was eroding democracy by concentrating “[power, resources and decision-making in an engineering elite](https://ethics.harvard.edu/files/center-for-ethics/files/aifailsus.jhdcarr_final_2.pdf?m=1651510742).”

What’s changed is the timeline for action. The sudden release of ChatGPT into the wild by tech companies spurred by competitive pressures and despite the technology’s profound limitations and known harms has:

1. Coalesced community concerns about current harms from AI, particularly for Australians for whom this period coincided with news from the [Robodebt Royal Commission](https://robodebt.royalcommission.gov.au/),
2. Drawn attention to the foreseeable and potential harms from untrustworthy AI about which experts has long been concerned
3. Reminded citizens of the failure of western governments to regulate so as to prevent or mitigate the harms from the last digital disruption, leading to continued violations of privacy, attention theft, and substitutions of screen time for human time resulting in the [public health hazard of loneliness](https://www.forbes.com/sites/bryanrobinson/2023/05/06/us-surgeon-general-cites-loneliness-as-serious-mental-health-hazard-in-new-report/?sh=55510dde1adc) and the undermining of young people’s [acceptance of their bodies](https://www.theguardian.com/technology/2021/sep/14/facebook-aware-instagram-harmful-effect-teenage-girls-leak-reveals), [mental health and resilience](https://greatergood.berkeley.edu/article/item/how_teens_today_are_different_from_past_generations)
4. Reminded Australians of the demonstrated insufficiency of self-regulation to manage current, foreseeable or future harms from technology

In mid-July 2023, the Federal Trade Commission in the US opened an investigation into OpenAI to determine whether ChatGPT has harmed consumers through its collection of data and its publication of false information on individuals. The agency’s chair [believes that](https://www.nytimes.com/2023/07/13/technology/chatgpt-investigation-ftc-openai.html?campaign_id=190&emc=edit_ufn_20230713&instance_id=97431&nl=from-the-times&regi_id=211749458&segment_id=139214&te=1&user_id=121f3bfed4381b48be220210929854b7) tech companies should be regulated while technologies are nascent, to avoid a repeat of the regulatory failures that characterised the last digital disruption. As [she wrote](https://www.nytimes.com/2023/05/03/opinion/ai-lina-khan-ftc-technology.html) in the New York Times earlier this month:

What began as a revolutionary set of technologies ended up concentrating enormous private power over key services and locking in business models that come at extraordinary cost to our privacy and security. The trajectory of the Web 2.0 era was not inevitable — it was instead shaped by a broad range of policy choices. And we now face another moment of choice. As the use of A.I. becomes more widespread, public officials have a responsibility to ensure this hard-learned history doesn’t repeat itself.

**Are there any further non-regulatory initiatives the Australian Government could implement to support responsible AI practices in Australia?**

We see the training of civil society, business and government leaders, to develop the judgement and courage required to give voice to their values, and to design their organisations in ways that support and encourage their staff to do the same. As the [World Economic Forum’s Ethics by Design report](https://www.weforum.org/whitepapers/ethics-by-design-an-organizational-approach-to-responsible-use-of-technology/) observes:

“While certain foundational ethical risks can be mitigated through the establishment of clear operational rules, many others require the capacity for ethical judgement and organisational factors that support translating that judgement into action. As technology is increasingly incorporated into the daily operations of companies…leaders must prepare their people to be aware of the ethical risks…[and to] equip them to make ethical choices…”

As recent ethical scandals in the [Banking and Finance industry](https://www.royalcommission.gov.au/banking), at [Price Waterhouse Coopers](https://www.theguardian.com/australia-news/2023/may/31/pwc-australia-scandal-what-actually-happened-and-will-it-be-fatal-for-the-advisory-firm#:~:text=PricewaterhouseCoopers'%20public%20slogan%20is%20to,to%20extend%20beyond%20national%20borders.) and [at the top of government departments](https://www.theguardian.com/australia-news/2023/jul/10/seven-public-servants-criticised-in-robodebt-report-as-agencies-consider-response) charged with delivering vital services to the most vulnerable Australians show, senior leaders and established institutions often struggle to recognise the ethical components of a decision, little less apply themselves to parsing and resolving them in ways that align with personal and organisational values.

Ethical leadership training can help, particularly training that balances the traditional focus on choosing correctly with a concern about the courage required to act. A focus on the “ethical blind spots” that can lead good people astray, and the “choice architecture” that can help organisations stay true to their values, is also critical and aligns well with the by-design focus of effective digital legislation. Doing the right thing, the thing that engenders trust, must be the default, or otherwise baked right in.

For some organisations, just including ethical considerations when decisions are made at board and senior levels - along with information about finance and legal advice, would be a change, one that could be achieved by a range of interventions - from checklists to ethics refresher courses - to keep ethics top of mind. Missions statements and the like also matter, while removing financial incentive structures is critical, as they sap the intrinsic motivation to do good and, as seen during the Banking Royal Commission, are often the source of unethical and unlawful conduct.

The more power technology gives us to act, and to act more powerfully, the more leaders and their organisations will be faced with hard ethical choices. How they make them, and what they decide, will impact the trust Australians have in AI. The Australian AI Commission should keep a list of reputable providers of Ethical AI Leadership Training that meets a set of competencies and standards.