# Response to the Supporting responsible AI: discussion paper

Katherine Biewer

**Foreword:**

In the following sections, I present my answers to questions 1, 2, and 20.

# 1. Do you agree with the definitions in this discussion paper? If not, what definitions do you prefer and why?

## Definitions are not dynamic

The definitions outlined in Figure 1 are both too specific in some respects and lack important detail in others. The definitions focus primarily on defining a set of “technologies” and “applications” that do not appropriately encapsulate the spectrum of capability, complexity and risk associated with AI.

For example, the definition of Machine Learning, while not incorrect, is oversimplified. Machine Learning is not merely about deriving patterns from training data; it is a subset of AI that uses statistics to give algorithms the ability to "learn" from data, thereby improving their performance on a specific task over time without being explicitly programmed to do so. It includes a variety of techniques such as Supervised Learning, Unsupervised Learning, Reinforcement Learning, etc., each with its own set of assumptions, strengths, and weaknesses. Reinforcement Learning, which I will address more robustly in a later response, is by its nature far scarier than other forms of Machine Learning as these systems are trained to “want” things. That is, they are not passive (i.e., require prompting for an output), but actively operate in their environment and pursue specific goals. The umbrella term ‘Machine Learning’ is not a sufficient enough term to distinguish the different capabilities associated with its various subsets.

Further, a distinction should be made between Machine Learning and Deep Learning. Traditional Machine Learning algorithms, such as tree-based methods and support vector machines, typically only work well on small to medium sized data sets. On the other hand, Deep Learning is a more complex type of Machine Learning that is inspired by the structure of the human brain and is particularly effective when dealing with large amounts of data. Deep Learning models use artificial neural networks with multiple layers (hence the term "deep") to model and understand complex patterns in datasets. These layers of neurons allow deep learning models to learn features automatically from the data, eliminating the need for manual feature extraction, which is often required in traditional Machine Learning. This clear distinction is integral when assessing the risks associated with AI.

The definitions are also seemingly short-sighted and rooted in the current state of AI, forgetting both its historical development and potential future advancements.

Historically, the field of AI has seen a transformation from rule-based systems, where the behaviour of the system is completely predetermined, to adaptive systems that can learn from data and improve over time. However, these simpler, low-risk systems

continue to operate today and should not be overlooked in any risk-based framework. Ignoring these simpler AI technologies may jeopardise the accurate risk classification for both lower-risk and higher-risk systems. An over-generalized risk assessment may understate the risks of complex AI while overstating that of simpler systems.

At the same time, the landscape of AI has greatly expanded to include a variety of more complex, higher-risk technologies. These advanced systems, some of which are currently operating within our society or being developed for future deployment, possess potential risks that far exceed those of simpler, rule-based AI systems. These could include autonomous weapons, development of novel biological agents or toxins, decision-making AI in critical infrastructure, or disinformation generators. These kinds of systems pose risks that are existential; they could lead to the end of civilization or lock humanity in an indefinite period of suffering.

The definitions provided in Figure 1, while generally accurate, fall short in capturing the intricate complexities and range of capabilities inherent to AI technologies. By oversimplifying concepts such as Machine Learning and not differentiating between its various subsets, the definitions fail to encapsulate the richness of these technologies. Moreover, the current definitions fail to anticipate the potential advancements in AI, and the consequential emergence of new subsets and applications. Therefore, the field would benefit from a set of definitions that are not only more nuanced and precise but also designed to evolve alongside the rapidly advancing landscape of AI. This is essential to foster a clearer understanding and dialogue about AI technologies in their full complexity, now and in the future.

## Recommendations for Improvement

1. **Model types:** Broaden definitions to encapsulate various methods and model types (i.e., break down Machine Learning into its subsets: supervised learning, unsupervised learning, deep learning, reinforcement learning, transformers, etc).

2. **Capabilities:** Include definitions about different AI capabilities, such as data processing abilities, temporal awareness (i.e., capacity for planning), narrow versus general applicability, multimodality (e.g., can it process both text and images?).

3. **Active vs Passive**: Distinguish between systems that are agentic (i.e., proactively interactive with their environment and pursue goals) and passive (i.e., require prompting or inputs for actions).

4. **Complexity:** Distinguish between simple (i.e., rule-based) AI and more complex (i.e., capacity for learning from data) AI systems.

5. **Risk:** Specifically acknowledge and define AI technology that carry potential existential risks, whilst also appropriately distinguishing systems that are low-risk.

# 2. 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?

## Acknowledge Existential Risk from AI

Australia must acknowledge the potentially existential risk posed by AI. Here, "existential risk" is a universal term encompassing all kinds of dystopian futures that don't just result in wiping humanity off the earth, but also the possibility of trapping us in for an endless period of suffering or sending humanity back into the dark ages.

There are two main narratives for how AI can contribute to the extinction (or prolonged suffering) of the human species:

- Misuse of highly capable AI.
- Unintended harm from highly capable

This existential risk is widely acknowledged by leading experts in AI, as highlighted by the following quotes in **Table 1** below.

**Table 1. Expert and Authoritative Perspectives on Existential Risk from AI**

| | |
|---|---|
| **Yoshua Bengio**<br><br>One of the "Godfathers" of AI | "I got around, like, 20 per cent probability that it turns out catastrophic." [1] |
| **Geoffery Hinton**<br><br>One of the "Godfathers" of AI | "I used to think it would be 30 to 50 years from now [before AI will supersede human level intelligence]. Now I think it's more likely to be five to 20." [2]<br><br>"How many cases do I know where something more intelligent is controlled by something less intelligent?" [3] |

[1] https://www.abc.net.au/news/2023-07-15/whats-your-pdoom-ai-researchers-worry-catastrophe/102591340

[2] https://www.wired.com/story/geoffrey-hinton-ai-chatgpt-dangers/

[3] https://www.theglobeandmail.com/business/article-i-hope-im-wrong-why-some-experts-see-doom-in-ai/

| | |
|---|---|
| **Sam Altman**<br><br>Founder and CEO of OpenAI | "AI will probably most likely lead to the end of the world, but in the meantime, there'll be great companies." [4] |
| **António Guterres**<br><br>UN Secretary General | "Alarm bells over [AI] are deafening, and they are loudest from the developers who designed it. These scientists and experts have called on the world to act, declaring AI an existential threat to humanity on a par with the risk of nuclear war." |

People often feel placated when they hear that the third “godfather” of AI, Yann Lecun, maintains that existential risk from AI is “preposterously ridiculous”[5]. Lecun often argues that "if you realise [AI is] not safe you just don't [build] it.” This argument is naive and fails to recognise that there are malicious agents out there who *intend* to build AI that is unsafe. This also neglects evidence of deceptive and emergent behaviours in AI systems.

The reduction of human extinction risk from AI has recently been declared a global priority, as articulated in the statement “Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war”. [6] This statement has been signed by numerous notable figures, including the heads of leading AI labs such as OpenAI, Google DeepMind, Anthropic, Stability AI, pivotal researchers and developers in the field like Geoffrey Hinton, Yoshua Bengio, and other influential individuals like Bill Gates.

The potential for catastrophic risk from AI, though not yet realised and uncertain, cannot be ignored. It is not acceptable for Australia to ignore an issue that experts and other countries have been actively engaging with.

## Suggested Regulatory Responses

**Strengthen Australia’s capacity for mitigating and responding to harmful, potentially catastrophic risks from AI.**

1. **Establish Conditional Usage Guidelines:** Develop a policy framework where AI usage is permitted conditionally, based on the public benefit. Disallow use of high-risk AI technologies that do not contribute to the benefit of society (i.e., intrusive advertising, automated social media bots, etc).

---

[4] https://futureoflife.org/ai/sam-altman-investing-in-ai-safety-research/?utm_source=Sailthru&utm_medium=email&utm_campaign=Future%20Perfect%205.24.23&utm_term=Future%20Perfect

[5] https://www.bbc.com/news/technology-65886125

[6] https://www.safe.ai/statement-on-ai-risk#open-letter

2. **Strengthen Regulatory Oversight of AI Development:** Empower governmental agencies with the authority to conduct inspections and audits pertaining to risk from AI development.

3. **Mandate Access to AI Models for Auditing:** Enforce a policy where AI labs are required to provide regulators with access to their AI models' weights and structures. This will allow for effective auditing and verification of the technology's safety, ethical standards, and potential impacts, before it is allowed to operate within Australia.

4. **Monitor indicators of AI development:** This includes both transparent and overt development via tracking research publications, patent applications, investment in AI, release of open-source or proprietary models, etc, and also covert development via tracking anomalous surges in data and power usage and acquisition of computing hardware.

5. **Strengthen Legal Liability:** Ensure AI labs cannot shift liability "downstream", and provide clear legal recourse for consumers and businesses when rights are violated due to AI-related harms.

6. **Implement Joint Culpability Scheme:** Given the opaque nature of AI systems, AI labs should hold liability for the consequences of their systems, with consumers and businesses sharing a part of the responsibility.

7. **Introduce Mandatory Reporting and Response Mechanisms:** AI labs should be required to actively monitor for and report any malicious use of their technologies. Additionally, they should have a plan in place to respond to and mitigate any harm caused by such misuse. This includes cooperating with law enforcement agencies and regulatory bodies as necessary.

**Strengthen Australia's National AI Security and Safety through Technological Literacy and Enhanced Research Initiatives**

1. **Improve Technological Literacy within Government:** Initiate educational programs to enhance understanding of AI and related technologies within the government and public service sectors, fostering informed decision-making.

2. **Promote Public Technological Literacy:** Implement campaigns and public education programs to increase understanding of AI, its potential risks, and mechanisms to mitigate them.

3. **Establish AI Safety Research Body:** Set up an Australian national technical laboratory to analyse, monitor, and ensure safety and interpretability of AI systems, similar to the UK's AI Sentinel[7] or Singapore's AI Verify Foundation[8].
4. **Boost AI Safety Research:** Enhance funding and support for AI safety research within Australian universities and institutions, encouraging innovation in AI safety.

---

7 https://www.institute.global/insights/politics-and-governance/new-national-purpose-ai-promises-world-leading-future-of-britain

8 https://aiverifyfoundation.sg/

# 20. Should a risk-based approach for responsible AI be a voluntary or self-regulation tool or be mandated through regulation?

## A Warning Against AI's Self-Regulation Drawn from the Deepwater Horizon Catastrophe

Much like the catastrophic Deepwater Horizon oil spill that laid bare the deadly shortcomings of self-regulation in a high-risk industry, we stand on the precipice of a potential disaster of equal or greater magnitude with artificial intelligence if self-regulation is allowed unchecked.

In the lead-up to one of the largest oil spills in history, regulatory mechanisms failed in two critical areas:

1. The Act governing development of oil and gas resources in the region: the Outer Continental Shelf Lands Act ('the OCSLA').
2. The overseeing agency responsible for implementing the Act: the Minerals Management Service ('the MMS').

The OCSLA did not contain robust, enforceable mandates for adequate environmental and safety standards. Instead, the Act left it to the discretion of the agency (the MMS) to determine the environmental and safety standards to be upheld by oil drilling operators. However, the MMS lacked appropriate funding, resources and expertise[9] to independently develop and implement robust, enforceable regulations, causing it to rely heavily on industry self-monitoring and adopt industry standards. This meant the industry had free rein to set the standards that were ultimately implemented to regulate them, with the agency (the MMS) set to verify these standards lacking both the domain knowledge and resources to provide meaningful oversight.

As a consequence, safety was overlooked by the industry, prioritising economic benefits with no incentive to innovate stricter safety measures and technologies, nor any sufficient authority or binding legislation to hold them accountable. This regulatory failure has been recognised as the most significant factor leading to the Deepwater Horizon incident.

In light of this, I will draw parallels between the regulatory failures preceding the Deepwater Horizon disaster and potential similar pitfalls that may emerge in AI governance, to make a case for:

---

9 https://www.ucsusa.org/resources/bp-and-other-companies-exploited-regulatory-agency-continue-negligent-offshore-drilling

1. The insufficiency of purely voluntary, self-regulatory measures in ensuring the safe development of AI.
2. The urgent need for robust and enforceable safety regulations.
3. The crucial role of a well-resourced, technically competent supervisory agency in upholding these regulations.

## Issues of Self Regulation in a High-Risk, High Reward Industry

Much like the AI industry, technological advancement in the oil industry at the time of the Deepwater Horizon disaster was rapid, driven by high economic returns and increased productivity. The governance of high-risk industries often grapples with the trade-off between economic progress and stringent regulatory oversight. The tragic Deepwater Horizon oil spill served as a potent reminder of what happens when this balance tips too far in favour of the former, with lax regulations and industry self-governance leading to catastrophic consequences.

### Self-Regulation Amid Perverse Incentives is a Bad Idea

Without any safety mandates in place to hold them accountable, the oil industry was allowed to set their standards for operation. These standards, rather unsurprisingly, were in favour of the industry's interests (i.e., to maximise oil extraction) instead of prioritising safety and environmental protection. No regulations were in place to force innovation of safer practices or effective responses to potential oil spills, despite the rapid growth in oil extraction capacities. Disturbingly, the MMS simply accepted assurances that a rig blowout was unlikely without any investigation and resigned the obligation of oil spill mitigation to “the limitations of available technology”.[10] This regulatory model ironically discourages the oil industry from pioneering technologies that can manage larger oil spills, as doing so would increase their clean-up responsibilities. In 2002, eight years prior to the oil spill disaster, the Coast Guard cautioned that without regulatory pressure, oil companies wouldn't create new technologies to prevent and respond to oil spills, even as their extraction capabilities grew [11]. Sound familiar?

This deficiency in adequate safety and response measures tragically manifested during the 87-day-long Deepwater Horizon oil spill, which resulted in the release of 3.19 million barrels of oil into the ocean[12]. Horrifyingly, responders at the forefront of the disaster were armed only with a response plan that was full of holes and painted “a picture of a

[10] https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1685606
[11] https://www.dco.uscg.mil/Portals/9/CG-5R/MER/MER%203/SERG_04_April_2019_FINAL.pdf
[12] https://www.fisheries.noaa.gov/news/deepwater-horizon-10-years-later-10-questions#:~:text=1)%20How%20large%20was%20the,of%20oil%20into%20the%20ocean.

company that was making it up as it went along, while telling regulators it had the full capability to deal with a major spill". [13]

The circumstances of AI development present worrying similarities. Like the oil industry, the development and deployment of AI technologies are currently largely in the hands of private enterprises, with minimal regulatory oversight. The financial incentives for creating increasingly sophisticated yet riskier AI systems are considerable, these systems are advancing so swiftly that even AI experts struggle to keep pace with their evolution [14]. Although AI capabilities have seen rapid growth, investment in AI safety research hasn't kept pace. This is evident at OpenAI, who have dedicated only 20% of their total computing capacity toward AI alignment research[15], a commitment that feels disproportionately low given the potential risks of advanced AI. Further, in June 2023, it was estimated that there were "only 300 people working on technical approaches to reducing existential risks from AI systems."[16]

There's a danger of repeating past mistakes by implementing a regulatory framework for AI that mirrors the one in place for the oil industry before the Deepwater Horizon disaster – a framework that relies heavily up industry self-regulation, favouring economic benefits and tethering the responsibility of safe AI development to the "limitations of available technology".

We must avoid repeating the same mistakes when governing AI at all costs. Although the oil spill was one of the most devastating environmental catastrophes, taking the lives of 11 people, the level of risk posed from AI is catastrophically greater.

It's crucial to implement enforceable, binding regulations that demand transparency, enforce strategies to mitigate risk, and hold AI labs accountable for their technologies throughout the whole lifecycle, from inception to deployment. We should demand robust evidence of safe development, subject to audit by adequately trained individuals, and set strict benchmarks. We must promote and enforce the innovation and adoption of increasingly safer strategies.

---

[13] https://www.nbcnews.com/id/wbna37599810

[14] https://futureoflife.org/podcast/roman-yampolskiy-on-the-uncontrollability-incomprehensibility-and-unexplainability-of-ai/

[15] https://openai.com/blog/introducing-superalignment?fbclid=IwAR1Cuugnm5dSCp64K9zv-umzPVXMasGwmfTCwtSGV1dnHP4MrDNrz3vUrbk

[16] https://80000hours.org/career-reviews/ai-safety-researcher/?utm_source=google&utm_medium=cpc&utm_campaign=80KMAR-AISafetyCareersBroad&utm_content=149150470961&utm_term=artificial%20intelligence%20safety&gclid=Cj0KCQjwiIOmBhDjARIsAP6YhSUVK3CKFEEksONG3W1baNJ3MR89it7sqIUsMkEfnPdeJUqFvP7s3DYaAr2eEALw_wcB#neglectedness

## Industry jobs are attracting all the talent with their competitive salaries, with limited funding for agencies

Technological advancements, such as the shift from onshore to deepwater drilling, introduced significant complexities and risks that MMS was inadequately equipped to handle[17]. Without sufficient technical expertise and resources, it was virtually impossible for the MMS to effectively supervise these activities, let alone enforce robust safety standards. This rendered them effectively useless. The Deepwater Horizon disaster may have been averted if the MMS had the funding to proactively anticipate and respond to emerging technology; develop independent safety regulations that don't rely on industry self-regulation; hire, train, and retain capable staff; and monitor and enforce safety effectively.

Similarly, AI technology development is rapid. The overwhelming majority of AI researchers currently work in privately owned AI laboratories, where competitive salaries[18] and significant rewards, such as academic prestige and a boosted ego, fuel the creation of increasingly advanced systems. "Today, roughly 70% of individuals with a PhD in artificial intelligence get jobs in private industry". [19]

To avoid the missteps of the MMS and effectively govern AI, Australia must establish a well-funded agency with clear authority and technical capacity to audit AI safety effectively. The agency's staff should possess extensive technical literacy in AI and have a proven technical background. The agency must provide incentives and opportunities for ongoing learning to keep pace with the AI field's swift evolution. Equally important, the remuneration for these roles should be competitive, ensuring the attraction and retention of top talent amidst the lucrative offers from private AI labs.

## Lack of a clear regulatory framework fosters weak regulation

The MMS was severely handicapped due to the absence of a clear and rigorous regulatory framework. The OCSLA lacked specific and enforceable mandates, relying on the discretion of the MMS to establish environmental and safety standards. This ultimately led to weak regulation that was ineffective and susceptible to corruption.

Examples of weak regulation include the MMS permitting oil rig operations to continue for up to two years while their oil spill response plans were still pending approval[20]. In another instance, the MMS granted an exemption to BP, the lessees of the Deepwater Horizon rig, from the requirements of the National Environmental Policy Act (NEPA).

---

[17] https://www.govinfo.gov/content/pkg/GPO-OILCOMMISSION/pdf/GPO-OILCOMMISSION.pdf
[18] https://www.nytimes.com/2018/04/19/technology/artificial-intelligence-salaries-openai.html
[19] https://mitsloan.mit.edu/ideas-made-to-matter/study-industry-now-dominates-ai-research
[20] http://www.noia.org/wp-content/uploads/2015/12/DOI-OCS-Safety-Oversight-Board-Report.pdf

This effectively excluded them from the obligation to undertake a comprehensive assessment of the environmental impacts of their proposed drilling activities[21].

Regulations governing AI should not only guide its development and deployment but also explicitly define the responsibilities and limitations of the overseeing regulatory agencies, preventing them from arbitrarily defining standards. Regulatory agencies need to be held accountable, ensuring they operate with transparency and in public interest. This prevents them from becoming self-regulating bodies that could potentially fall into complacency, corruption, or overlook future-oriented risks associated with AI's vast potential.

## Conclusions

In conclusion, while a risk-based approach to AI governance is necessary, it should not be left to voluntary compliance or self-regulation. Given the potentially vast impacts of AI, a risk-based approach should be enforced through well-crafted, flexible, and comprehensive regulations. These regulations should encourage transparency, accountability, and ongoing innovations in safety and risk mitigation.

---

21 https://www.ucsusa.org/resources/bp-and-other-companies-exploited-regulatory-agency-continue-negligent-offshore-drilling