



A I S M

**AI SAFETY
MELBOURNE**

**SUPPORTING AN
EVIDENCE-BASED
APPROACH TO SAFE
AND RESPONSIBLE AI**



AI Safety Melbourne: Response to the Safe and Responsible AI Discussion Paper

Part 1: Introduction to AI Safety and our work	3
Personal Foreword from Justin Olive, AISM Coordinator:	3
Opening comments on the discussion paper:	4
1.2 Definitions:	4
2.1 Opportunities:	4
2.2 Challenges:	4
Part 2: The AI Landscape: July 2023 Update	5
2.1 Introduction to AI systems	5
2.2 Busting myths about AI	6
Table 1: Viewpoints of regular AI researchers compared to those with a track record of researching AGI or making significant contributions towards it.	7
2.3 Developing AI Systems:	8
2.4 Near-term future of AI	9
2.5 Defining AI Systems:	9
2.6 Access to AI Systems	10
2.7 Reasoning about AI as a structural force	11
2.8 Introduction to AI risks	13
2.9 New risks	13
2.10 Amplifications of existing risks	17
2.11: Technical Challenges for AI Regulation	18
Part 3: Recommendations	21
Recommendation 1:	21
Create distinct strategies for narrow AI vs general purpose AI	21
Recommendation 2:	21
Adopt four key principles for managing general purpose AI	21
2.1 Monitor AI Deployment:	22
2.2 Control Supply & Distribution Channels:	23
2.3 Defining Low-risk, high-value deployment:	24
2.4 Invest AI safety	25
Recommendation 3:	29
Implement a General-Purpose AI Risk Management Strategy	29
Part 1: Risk Assessment	30
Part 2: Precursor Systems	30
Part 3: Deployment of Advanced Systems	32
Part 4: Use of Advanced Systems	34



Part 1: Introduction to AI Safety and our work

AI safety is an area of research that aims to uncover how deep-learning based systems can be engineered to be more:

- **Controllable:** i.e. their actions and outputs can be directed with specific detail
- **Transparent:** so that behaviours and traits of a systems can be explained
- **Predictable:** meaning that behaviours are consistent, and robust against anomalies.

AI Safety Melbourne (AISM) is a community group for individuals who are interested in contributing to the field of AI safety. We have run events with expert presenters and guests from a range of prominent organisations:

- **IT & Data professionals** (Atlassian, Microsoft, ANZ, Accenture, Decoded.ai)
- **Researchers and lecturers** (University of Melbourne, Monash University)
- **Public servants** (Department of Defence, Department of Treasury and Finance, Department of Education)

Personal Foreword from Justin Olive, AISM Coordinator:

I'm very concerned by the trend of seeing young, high-quality researchers leave Australia. Of the presenters we've had at our recent events, two are leaving for the UK to pursue better opportunities in AI Safety.

The most common barrier to organising events is finding a date when the presenter is in the country, because they're often travelling to overseas conferences, or visiting overseas organisations.

One of the most common discussions I hear at our events is whether or not to move to London or San Francisco; Australia's AI ecosystem is essentially seen as irrelevant.

I've had a software engineer in the AISM community explain to me how they made an AI-related presentation to researchers from CSIRO's Data61, and found that they didn't even know the techniques shown were possible.

When seeking feedback on this submission, some members simply said that Australian regulation doesn't matter enough to worry about it, because we're not the ones developing advanced capabilities or doing AI safety research.

As someone who is trying to build up Australia's AI ecosystem and make it globally impactful and beneficial, experiences such as these can be frustrating. I hope these observations help colour the current state of AI in Australia, and how there is a dire need for investment in the Australian AI safety ecosystem.

Kind regards,

Justin

Opening comments on the discussion paper:

1.2 Definitions:

The approach to structuring the definitions does not reflect a mature understanding of machine learning or AI. Improvement is needed, including a discussion of:

- General AI vs Narrow AI
- Autonomous vs Static systems
- Main approaches to ML (e.g. supervised vs reinforcement learning)

Further detail in Part 2 (2.1 - 2.6)

2.1 Opportunities:

- It is promising to see reference to the speed of research advancements, the uncertainty of future progress, and the “rapid emergence of open-source systems”

2.2 Challenges:

Positive Feedback:

- Discussion of transparency is crucial, although it would be ideal to emphasise that AI explainability is the foundation of transparency.
- Fantastic to see that the authors identifying the trend of proprietary datasets leading to “economically powerful organisations developing and deploying sophisticated AI”

Criticisms:

- Did not discuss structural risks, catastrophic risks, or specific risks from deceptive and power-seeking systems
- Did not address the implications of systems that will likely come with emerging applications in autonomous systems based on foundation models.
- Did not address most of the actual challenges of regulating AI systems, such as replicability, autonomy, misalignment in deep learning systems, etc.

Further detail in Part 2 (2.7 - 2.11)

3: Domestic and international landscape

- Given that AI is a constantly evolving issue, it matters less what other countries have already implemented, and more about what expert organisations, such as the Centre for AI Governance, are recommending. Australia should be focusing on being an agile leader in AI governance.
- Does not appear to address the likely and desirable scenario that an international AI regulator will be formed which (Australia should advocate for). For information on this, see:
 - Ho et al. (2023) International Institutions for Advanced AI;
<https://arxiv.org/abs/2307.04699>

4: Managing potential risk:

- Promising to see that the Government is considering options such as new AI laws,, Legislated technical requirements, Mandatory registers of AI, Mandatory Certification. These align with current recommendations from AI governance researchers (see Part 3 for how these should be implemented by the Australian Government)
- The draft risk management framework in Box 4 is very concerning. Among other things, it demonstrates an understanding of AI risks that is deeply insufficient. The Government's understanding of AI risks needs to be informed by research from AI safety and Governance experts. See Part 2 for a summary, although additional resources include:
 - Anderljung et al. (2023) *Frontier AI Regulation: Managing Emerging Risks to Public Safety*; <https://arxiv.org/pdf/2307.03718.pdf>
 - Hendrycks et al. (2023) *An Overview of Catastrophic AI Risks*; <https://arxiv.org/pdf/2306.12001.pdf>
 - Hendrycks et al. (2022) *Unsolved Problems in ML Safety*; <https://arxiv.org/abs/2109.13916>
 - Goldstein et al. (2023) *Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations*; <https://arxiv.org/pdf/2301.04246.pdf>
 - Anderljung & Hazell (2023) *Protecting Society from AI Misuse: When are Restrictions on Capabilities Warranted?* <https://arxiv.org/pdf/2303.09377.pdf>
- Promising to see changes to assurance infrastructure being raised, as these will be key. In particular, the following was a salient point to raise:
 - “building explainability into AI systems that could incorporate by-design considerations amongst other things to support greater transparency.”

Overall, the Discussion Paper indicates an encouragingly open-minded framing of AI regulation. However, it does seem to indicate that the Government might be underestimating the magnitude of the issues that are on the horizon. A risk-based approach is essential, but this is only useful when paired with a well-informed understanding of the risks that AI safety researchers have highlighted.

Nevertheless, we're confident that the Australian Government will have the maturity and agility to align its regulatory response with the current realities of the AI landscape.

Part 2: The AI Landscape: July 2023 Update

We have observed that Governments around the world tend to only show awareness of themes 2-3 years after they first emerge. As a group who are highly engaged in the forefront of AI, we feel a duty to provide the Australian Government with an up-to-date overview of the current landscape and future directions.

Accordingly, the following section is an update to clarify how the Government should be thinking about AI in 2023, and it draws upon conversations held among AISM's diverse range of experts, as well as high-quality research into AI Governance. The definitions and evidence provided are essential for understanding why the recommendations introduced in Part 3 are valuable, and the details of how they should be implemented.

2.1 Introduction to AI systems

Artificial Intelligence (AI) is a general class of technologies that combine three key factors to achieve impressive performance on nontrivial tasks. These key factors are¹:

- Algorithms: particularly neural networks
- Large Datasets: gigabytes to petabytes
- High-performance computing (e.g. GPUs)

AI systems are generally composed of one or more **machine learning** (ML) models. For example, a general-purpose AI system such as ChatGPT can be composed of GPT-4 (a general-purpose ML model), as well as other narrow models involved in filtering and monitoring.

A form of ML models known as neural networks (i.e. deep learning) are particularly relevant to AI governance, and should be thought of as a distinct category compared to other common methods. This is because neural networks learn in a way that is fundamentally more “human-like”; i.e. their unique mathematical properties offer advantages to generalisability and scalability compared to other statistical models^{2, 3}.

In many cases, the architecture of a neural network is not inherently complicated; many can be described as massive blocks of numbers (known as **parameters**) that produce an output by performing a series of calculations on an input. The sophistication lies in how these blocks of numbers are shaped during **training** (i.e. where the neural network “learns” from the data) - this is a complex, computationally-expensive process.

¹ Sevilla et al. (2022) *Compute Trends Across Three Eras of Machine Learning*; <https://arxiv.org/pdf/2202.05924.pdf>

² Murfet et al. (2023) *Deep Learning is Singular, and That's Good*; <https://arxiv.org/abs/2010.11560>

³ Nagayasu & Watanabe (2023) *Bayesian Free Energy of Deep ReLU Neural Network in Overparametrized Cases*; <https://arxiv.org/pdf/2303.15739.pdf>

2.2 Busting myths about AI

- **“AI is just advanced statistics”**: AI systems require significant resources and engineering efforts to develop, which involves statistical concepts as one ingredient. If something is “just advanced statistics”, then it is probably not AI.
- **“AI is everywhere, we use AI every day”**: media coverage around AI often conflates two very different forms of technology⁴. This confusion is mostly caused by the rampant use of AI as a marketing term for any software that involves machine learning. Applications such as pattern recognition or predictive analytics can be conceptualised as **“narrow AI”**, although from a regulatory perspective, it is ideal to treat these systems more like “software features” than a distinct class of technology. In today’s AI landscape, it is becoming more important to focus on general purpose AI; this is something the EU is now recognising⁵.
- **“AI is just maths; it isn’t special/scary, and it will never be able to do X”**: statements of this kind are common in the media. Unfortunately, these misleading quotes are often taken from AI experts who are either focused on narrow AI⁶, or have expertise in areas other than deep learning⁷.

When reviewing claims made by an “AI expert”, it is crucial to examine the relevant expertise behind these beliefs. From an intuitive perspective, this is similar to how a guitar teacher cannot provide expert advice on how to play piano; there are underlying similarities, but the details are different and relevant experience is key.

Unfortunately, there are comparatively very few credible experts in artificial general intelligence. Importantly, their perspectives also tend to differ greatly compared to regular AI researchers. For a range of examples that illustrate this point, see Table 1.

- **“If AI is so smart/dangerous, why can’t it do X”**: Humans evolved to have a set of very specialised abilities, including verbal and non-verbal communication, navigating uncertainty, and spatial reasoning. Like a sprinter who calls a marathon runner “slow”, we’re missing out on the full picture when we focus on narrow, biased criteria.

Our tendency to do this⁸ means that we struggle to understand the implications of *Moravec’s Paradox**, a principle that describes how humans and machines tend to excel at opposite types of tasks⁹. For example, it is relatively straightforward for computers to solve complex logic or optimisation problems, like playing chess. This means that, as AI systems approach human-capabilities in tasks that humans already specialise in, they’re actually far surpassing humans in overall intellectual capabilities.

*AI experts often joke that stacking a dish-washer will be the last task to be automated by AI, which is an intuitive illustration of Moravec’s Paradox.

⁴ Toby Walsh (2022) *Is artificial intelligence a friend or foe? Or is that a question for Siri?*; SMH

⁵ White & Evans (2023) *The AI Act – A step closer to the first law on Artificial Intelligence*; dataprotectionreport.com

⁶ Oren Etzioni (2016) *Deep Learning Isn’t a Dangerous Magic Genie. It’s Just Math*; Wired Magazine

⁷ Toby Walsh (2017) *Elon Musk is wrong. The AI singularity won’t kill us all*; Wired Magazine

⁸ Field, M (2023) *‘Stupid’ AI won’t eradicate humanity, says Sir Nick Clegg*; [Telegraph UK](https://www.telegraph.co.uk)

⁹ Thakur, V (2022) *What is Moravec’s Paradox?*; [Science ABC](https://www.scienceabc.com)

Table 1: Viewpoints of regular AI researchers compared to those with a track record of researching AGI or making significant contributions towards it.

Regular AI Expert	AGI-focused Deep Learning Researcher
Data61 Director John Whittle warns against “speculating about superintelligence”, saying there is “no evidence that such an imminent threat exists”. ¹⁰	Yann LeCunn, Facebook Chief AI Scientist and Turing Award Winner, suggests we will have human-level AGI in 10-15 years, and has proposed a detailed roadmap for arriving there. ¹¹
An Australian AI PhD student published in <i>The Conversation</i> writes: “... catastrophic AGI scenarios depend on premises I find implausible” ¹²	Geoffrey Hinton, a pioneer in deep learning, says that superhuman AI will probably arise in the next 5-20 years, and that it poses catastrophic risks ¹³ .
UNSW Chief AI Scientist Toby Walsh says: “Evolution has equipped us to want things, but machines don’t have any wants ... it’s not sitting there thinking, You know what? I want to take over the universe.” ¹⁴	Marcus Hutter, a DeepMind Senior Scientist researching AGI, writes: “... deploying a sufficiently advanced reinforcement learning agent would likely be incompatible with the continued survival of humanity.” ¹⁵
Professor Melanie Mitchell is a renowned leader in an approach to AI called “genetic algorithms”. She has been vocal in expressing great doubts that human-level AI will arise in the near-future, or that catastrophic risks are possible. ¹⁶	Jürgen Schmidhuber is a deep-learning pioneer and inventor of the LSTM network. He says that the intense competition within ecosystems of advanced AI will be “beyond our imagination”, and that it’s inevitable we will be outcompeted and replaced. ¹⁷
Several renowned European ML researchers in areas such as causal inference and optimisation algorithms have put forward a joint statement urging the public to focus on “narrow AI” rather than speculate about “hypothetical” scenarios such as AGI or superintelligence ¹⁸ .	Demis Hassabis, CEO of DeepMind, says human-level AGI is “maybe within a decade away”. ¹⁹
	Yoshua Bengio, an award-winning pioneer in deep learning, says that superintelligence is 5 to 20 years away, and carries the potential of catastrophic risks. ²⁰
Oren Etzioni is CEO of the Allen Institute for AI, and an expert in web search. He has historically pushed back against speculation about “superintelligence” or notions that it might pose catastrophic risks. ²¹	Paul Christiano, an eminent AI Safety researcher who helped develop GPT-4, says there is a 46% chance that “humanity has somehow irreversibly messed up our future within 10 years of building powerful AI”

¹⁰ Whittle et al. (2023) *Hype or fear: the AI debate examined*; [CSIRO Website](#)

¹¹ Heikkilä & Heaven (2022) Yann LeCun has a bold new vision for the future of AI; [MIT Technology Review](#)

¹² Bennett, M (2023) *No, AI probably won’t kill us all – and there’s more to this fear campaign than meets the eye*; The Conversation AU

¹³ Hinton, G (2023) *Why the godfather of AI fears for humanity*; [The Guardian](#)

¹⁴ Ange Lavoipierre (2023) AI’s dark in-joke; ABC News

¹⁵ Cohen & Hutter (2022) The danger of advanced artificial intelligence controlling its own feedback; [The Conversation](#)

¹⁶ Transcript (2023) Is AI an existential threat? Yann LeCun, Max Tegmark, Melanie Mitchell, and Yoshua Bengio make their case; [The Hub](#)

¹⁷ Schmidhuber, J (2017) *Falling Walls: The Past, Present and Future of Artificial Intelligence*. Scientific American

¹⁸ Oliver et al (2023) *Let’s focus on AI’s tangible risks rather than speculating about its potential to pose an existential threat*; [The Conversation](#)

¹⁹ Demis, Hassabis (2023) *DeepMind boss says human-level AI is just a few years away*; [The Independent](#)

²⁰ Bengio, Y. (2023) *FAQ on Catastrophic AI Risks*; <https://yoshuabengio.org/2023/06/24/faq-on-catastrophic-ai-risks/>

²¹ Etzioni, O (2016) *No, the Experts Don’t Think Superintelligent AI is a Threat to Humanity*; [MIT Tech Review](#)

2.3 Developing AI Systems:

There are three general approaches to training models relevant to AI systems:

1. **Supervised learning:** this is where each data point is labelled, either by a human or some automated process, and the ML model must learn to guess the label correctly. Example: image classification.
2. **Self-supervised learning:** this is when the model learns the structure of the data by using it both as the input and the label. Example: next word prediction.
3. **Reinforcement learning:** this is where the model learns by interacting within an environment and receiving rewards. In this case, the potential future reward for taking a given action is the label a model is trying to predict. Example: chess algorithm

For AI systems to achieve state-of-the-art performance on difficult tasks such as question answering and code generation, it is common to combine these training methods²²:

- This generally begins with a “**pre-training**” stage that uses self-supervised learning; this is where the model gains prior knowledge about relationships within the data.
- This is then followed by a “**fine-tuning**” process, which involves various forms of supervised learning and/or reinforcement learning. This shapes the raw pre-trained model to meet specific performance criteria by showing it specific examples of how to behave, or providing feedback on outputs.

After the training process is complete, the performance of a model on any given task depends on factors such as:

- The amount of training²³, and quality or size of dataset²⁴
- The size of the model (large models learning quicker)²⁵
- The type of fine-tuning processes used²⁶
- The prompt or environment used to operate the model (e.g. web interfaces, code executors, prompt optimisation)²⁷

2.4 Near-term future of AI

In order to overcome limitations on what a single model can achieve, AI developers have transitioned toward AI systems which chain together multiple models (e.g. the popular *LangChain* tool)²⁸. These **multi-model AI systems** can be constructed by duplicating a single

²² Wang et al. (2023) *Interactive Natural Language Processing*; <https://arxiv.org/pdf/2305.13246.pdf>

²³ Muennighoff et al. (2023) *Scaling Data-Constrained Language Models*; <https://arxiv.org/abs/2305.16264>

²⁴ Longpre et al. (2023) *A Pretrainer's Guide to Training Data: Measuring the Effects of Data Age, Domain Coverage, Quality, & Toxicity*; <https://arxiv.org/abs/2305.13169>

²⁵ Hoffmann et al. (2022) *Training Compute-Optimal Large Language Models*; <https://arxiv.org/pdf/2203.15556.pdf>

²⁶ Wu et al. (2023) *Fine-Grained Human Feedback Gives Better Rewards for Language Model Training*; <https://arxiv.org/abs/2306.01693>

²⁷ Shinn et al (2023) *Reflexion: Language Agents with Verbal Reinforcement Learning*; <https://arxiv.org/abs/2303.11366>

²⁸ Chase, H (2023) *LangChain*; <https://perma.cc/U2V6-AL7V>.

model and assigning different roles to each version, or by constructing a network of different models with complementary capabilities²⁹.

AI systems are also being integrated with knowledge databases and memory systems; these assist the system in storing or retrieving useful information, and more recently, specific advice or skills that it has learned from previous tasks.

The latter is an example of how AI systems are being trained to act **autonomously** as “agents” which are capable of learning over time to pursue sophisticated goals. These AI systems have risen to prominence in 2023, and have shown impressive preliminary results in interacting with digital environments³⁰.

Methods for training these autonomous, multi-model AI systems are an untapped research area with extraordinary potential for impact^{31, 32}. The emergence of foundation models the next decade will be spent harvesting this unprecedented level of low-hanging fruit in the area of general-purpose AI agents.

At this stage, the specific outcomes from this research are still uncertain. However, many reputable sources estimate that human-level AGI will arise in the next 5-20 years, meaning there is roughly a 50% chance AI systems will exceed human capabilities by 2035 (see Table 1 for specific predictions).

Key point: *regulators must be prepared for the near-term emergence of human-level AI systems which can autonomously plan and complete any tasks that involve reasoning about information and interacting with a digital environments³³.*

2.5 Defining AI Systems:

AI systems which approach or exceed human reasoning capabilities can be simply referred to as **advanced AI systems**³⁴. One of the first steps to prepare for these systems is to specify the relevant criteria in more precise terms. As such, we can categorise “advanced AI systems” as those which:

1. Are able to complete a diverse range of reasoning tasks with human-level performance; and
2. Are able to navigate complex information environments and act within those environments autonomously; and
3. Are able to form sophisticated plans and reason about the consequences of actions

Note: given the current pace of AI research relevant to current capabilities, it is possible that various forms of advanced AI systems will begin to emerge in the next 1-3 years.

²⁹ Yang et al. (2023) *Foundation Models for Decision Making: Problems, Methods, and Opportunities*; <https://arxiv.org/pdf/2303.04129.pdf>

³⁰ Nvidia: Wang et al. (2023) *Voyager: An Open-Ended Embodied Agent with Large Language Models*; Demonstration available: <https://voyager.minedojo.org/>

³¹ Hu et al. (2023) *Enabling Intelligent Interactions between an Agent and an LLM: A Reinforcement Learning Approach*; <https://arxiv.org/pdf/2306.03604.pdf>

³² Boiko et al. (2023) *Emergent autonomous scientific research capabilities of large language models*; <https://arxiv.org/abs/2304.05332>

³³ Kaddour et al. (2023) *Challenges and applications for LLMs*; <https://arxiv.org/pdf/2307.10169.pdf>

³⁴ Ho et al. (2023) *International Institutions for Advanced AI*; <https://arxiv.org/abs/2307.04699>

Another necessary definition for regulators which we introduce is **precursor systems**. These are models or AI systems which:

- Can be used to construct an advanced system within a set resource budget (e.g. <\$50,000 of labour hours and compute resources)
- Are central to the reasoning capabilities of the AI system being built (i.e. not tasked with perception or translation).

2.6 Access to AI Systems

In the context of AI regulation, it is also important to distinguish between AI systems in terms of how they are made available to the public. This is mostly captured in the difference between closed-source vs open-source AI systems, both of which have distinct advantages and risks.

Closed-source AI systems are those which have been developed by a company as a proprietary service, and can only be accessed via an API. An example is a system such as ChatGPT, or OpenAI's GPT-4 API which can be used to build products.

- The advantage of closed-source systems is that they offer more certainty over how it was developed, what it is being used for, and who is using it.
- The disadvantage of closed-source systems is that it is more difficult for the research community to ascertain the safety and reliability of the system.

Open-source AI systems are those which have been released to the public. I.e. the underlying "parameters" of the model can be downloaded so that the results of the original training process can be perfectly replicated.

- The advantage of open-source systems is that they offer a way for AI safety researchers to perform detailed examinations and tests on models which would otherwise cost thousands or millions of dollars to train. This helps uncover ways in which systems can be made safer or more explainable.
- The disadvantage of open-source systems is that they can be modified and deployed by anyone, with limited means of monitoring or restricting their use. This greatly lowers the barrier for actors who aim to use AI systems for malicious purposes.

Note: These conceptions of open or closed source are relevant in the current technology landscape, but regulators must be ready to monitor for changes and adapt as the field of AI advances. Examples of such advancements could include:

- Substantial advancements in AI systems that leverage information retrieval systems may trigger the emergence of open-source or proprietary knowledge bases that significantly influence the frontier of AI capabilities³⁵.
- As the use of meta-architectures (e.g. AutoGPT³⁶) matures, specific designs may become a key determining factor in the capabilities that a user can deploy.

³⁵ Jiang et al. (2023) *Active Retrieval Augmented Generation*; <https://arxiv.org/abs/2305.06983>

³⁶ Significant Gravitas (2023) *Auto-GPT: An Autonomous GPT-4 Experiment*; <https://perma.cc/2TT2-VQE8>.

2.7 Reasoning about AI as a structural force

The near-term potential impacts of advanced AI are commonly likened to ubiquitous technologies such as the internet or smart-phones³⁷. On the surface, this seems like a reassuring and sensible comparison.

However, the intense geopolitical tensions³⁸, regional monopolies³⁹, and arms-race dynamics⁴⁰ that characterise the current landscape indicate that advanced AI will be more analogous to powerful economic and strategic assets such as nuclear weapons and oil⁴¹, which together have played a large part in defining the previous century.

Similar to oil and nuclear weapons, the allure of advanced AI makes it irresistible to those who stand to gain from it. Yet events such as climate change and the Cuban Missile Crisis have taught us that things can go terribly wrong when powerful **structural forces** emerge⁴².

In the next section, we provide an examination of the political and scientific issues that have characterised crude oil. However, first it is important to make it clear why these are relevant to AI.

Crude oil fundamentally changed the limits of the economy. For example, it changed the type of technologies that were possible (introducing aeroplanes and satellites), as well as the scale and speed at which activities could be performed (e.g. single passenger travel).

AI is crude oil for the information age; it powers machines to do things that were previously not possible. Rather than combustion engines powering transport, it is computers powering information processing and decision making. A common mistake made when assessing the potential impact of this change is by using today's systems as the baseline, and then imagining how they might improve.

In fact, it is more useful to look at these changes from a more fundamental level.

For example, within a minute, GPT-4 can both process and respond to a 15,000-word document at a cost of <\$0.30 AUD⁴³. This means that an autonomous system based on GPT-4 can process at least 900,000 words of information per hour, at a cost of around \$18.

By comparison, humans can read approximately ~250 words per minute⁴⁴ and compose text at around ~19 words per minute⁴⁵. For comparative purposes, we will say that an average employee can process AND respond to 250 words per minute, or 15,000 words per hour, and that they can perform this for the minimum full-time wage of \$23.23 per hour.

³⁷ Dillian, J (2023) *AI Is the Next Internet... Expect Weird Things*;

<https://www.advisorperspectives.com/commentaries/2023/05/04/ai-is-the-next-internet-expect-weird-things>

³⁸ Cuellar & Sheehan (2023) *AI is Winning the AI Race*;

<https://foreignpolicy.com/2023/06/19/us-china-ai-race-regulation-artificial-intelligence/>

³⁹ Morgan, T (2023) *The Highly Profitable Chip Making Monopoly Called TSMC*;

<https://www.nextplatform.com/2023/01/12/the-highly-profitable-chip-making-monopoly-called-tsmc/>

⁴⁰ Stokel-Walker (2023) *TechScape: Google and Microsoft are in an AI arms race – who wins could change how we use the internet*; *The Guardian*

⁴¹ Ord, T (2022) *Lessons from the Development of the Atomic Bomb*;

⁴² Ord, T (2020). *The Precipice: Existential Risk and the Future of Humanity*.

⁴³ <https://help.openai.com/en/articles/7127956-how-much-does-gpt-4-cost>

⁴⁴ <https://www.sciencedirect.com/science/article/abs/pii/S0749596X19300786>

⁴⁵ Karat et al (1999) Patterns of entry and correction in large vocabulary continuous speech recognition systems; <https://dl.acm.org/doi/10.1145/302979.303160>

Despite this very favourable comparison, we still end up with the following:

- Automated Systems 50,000 words per dollar
- Human employee: 646 words per dollar

Statistics such as these begin to show why advanced AI will be a powerful structural force.

Oil - A case study in structural forces

Scientific challenges began in 1896, when Swedish scientist Svante Arrhenius first proposed that emissions from fossil fuels would lead to global warming. Arrhenius' findings were heavily criticised and quickly forgotten; the assertion that human activities could influence the atmosphere was seen as unrealistic, hypothetical and far-fetched⁴⁶. ([Table 1](#) makes it clear why this is relevant to AI)

While this response may now seem ignorant or negligent, in 1896, fossil fuel usage was still quite minimal by today's standards; catastrophic global warming would have been thousands of years away had the historical trends continued. Nevertheless, if scientists at the time had responded to these sensational claims with caution or curiosity rather than disbelief, society could have pursued the electrification of transport many decades earlier⁴⁷.

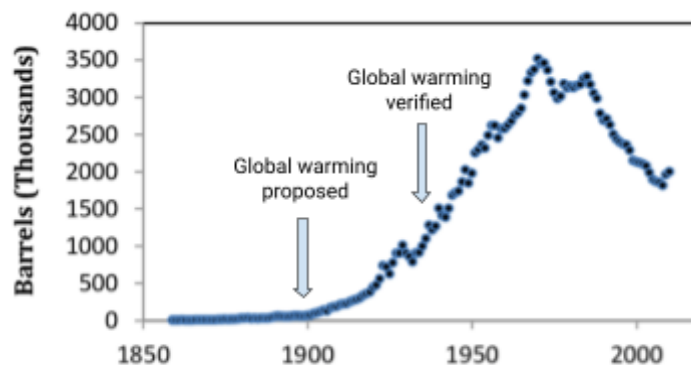


Figure 1: the world's largest producer of crude oil - USA production since 1859

In the 1940's, Arrhenius's proposal was proven correct. By this point, oil had become the bedrock of modern society, fuelling military and civilian transport networks, as well as becoming an indispensable manufacturing material. Alongside these benefits, oil also became the catalyst for significant economic and political turbulence, eventually earning it the pejorative title "*the Devil's excrement*". Some examples include:

- The emergence of violent petrostates such as Russia, Libya and Iraq⁴⁸
- Various invasions and wars, such as the Gulf War, and WWII conflicts between Germany-Russia⁴⁹ and Japan-USA⁵⁰
- Devastating accidents such as the 2010 BP oil spill, which caused one of the greatest environmental disasters of all time, and cost \$65 billion in remediation⁵¹.

⁴⁶ Maslin, M (2004) *Global Warming, a very short introduction*; Oxford Press via [Lenntech](#)

⁴⁷ Biography (2019) *Svante Arrhenius, the Man Who Foresaw Climate Change*; [OpenMind BBVA](#)

⁴⁸ Naim, M (2009) *The Devil's Excrement*; <https://foreignpolicy.com/2009/08/22/the-devils-excrement/>

⁴⁹ Encyclopedia Britannica; *The Germans' summer offensive in southern Russia, 1942*

⁵⁰ Yergin, D (1991) *Blood and Oil: Why Japan Attacked Pearl Harbour*

⁵¹ Vaughan, A (2022) *Deepwater Horizon oil spill did no harm to BP's long-term share value*; [newscientist.com](#)

Such occurrences illustrate that structural forces can become harmful by their very nature, in part due to the power they provide and the desperation they provoke. Handling such forces is rarely successful, and it is important to study rare instances where this is the case.

A classic example is Norway, who have managed their crude oil resources with such success that even they deem it to be a lucky turn of events. In truth, the only “luck” involved was discovering their oil reserves late enough to have already witnessed the wreckage that the gas industry inflicted on the Dutch economy in the decade prior⁵².

The true key to success was Norway’s public sector. Unlike other nations, whose political leaders and state institutions were subsumed by the oil industry, their regulatory response was proactive, cautious, and patient; within 2 years of discovering oil, they passed strict regulations to control its influence on their economy and political structures. This distinctive maturity is exemplified in a quote from Norway’s prime minister in 1975:

*“Professors and so-called experts from other countries give us advice to speed up oil production. (But) we don’t want it. **The point is to be sensible and careful**”.*

Because of their pro-regulatory approach, Norway has been able to take full advantage of their oil and gas industry, and now possess the world’s largest sovereign wealth fund, valued at \$2.0 trillion AUD.

2.8 Introduction to AI risks

The case study of Norway’s natural resources policies demonstrates how regulation channels powerful opportunities, rather than suppressing them. This works because powerful opportunities only yield benefits when the risks can be managed.

In the context of advanced AI, Australia must treat risk mitigation as the first priority, with potential benefits only considered as part of a long-term strategy for serving the public good. Just as the Norwegian bureaucrats learned from the failures of other oil states, the Australian Government must pay close attention to the warnings from AI safety researchers. Although the true risks from advanced AI will be difficult to predict, research has uncovered a number of very probable risks; these are generally framed as belonging to three classes, which have been explained below using the example of automobiles:

Risk Description	Automobile Example:
Misuse risks: where negligence or malicious intentions causes harm	Causing an accident after running a red light
Accident risks: where the AI system causes harm for unforeseen reasons	Causing an accident because the tail-lights stopped working.
Structural risks: where widespread deployment or availability contributes to negative downstream consequences	Air pollution from car emissions causes someone to develop cancer

⁵² Sing, S (2017) *The devil’s excrement: how Norway warded off the oil curse*; <https://www.livemint.com>

Note: When reviewing these risks, it is important to recognise that many will seem abstract and speculative; a natural response is disbelief or indifference. This is part of human nature, and it is similar to the feeling that someone from 1896 would have if they were told about oil causing issues like blitzkrieg warfare, sea-level rise or microplastics.

2.9 New risks

Category 1: Rogue AI: these are instances in which a dangerous and/or deceptive AI system is deployed:

1.1 Misaligned AI: current evidence about the behaviour of autonomous AI systems shows they are surprisingly difficult to monitor and control (often called the “alignment problem”)⁵³. This partly stems from the difficulty in specifying instructions or rewards so that an AI system pursues objectives in the ways that are intended.

Humans have evolved empathy and theory-of-mind reasoning to accurately guess intentions based on incomplete instructions. Unfortunately, it is very difficult to teach AI systems this ability. They’re also very good at finding effective tactics that humans would normally consider to be “cheating” or “deception”⁵⁴. Researchers are concerned that when we scale the capabilities of AI systems, we are also scaling their ability to deceive us⁵⁵.

Example - Misaligned AI: an AI agent trained to play the arcade game *Qbert* learned that it could maximise points by repeatedly committing suicide in order to defeat its enemies, while another found a pattern of movement that caused the game to glitch, immediately yielding a huge amount of points⁵⁶.

Game environments are ideal for testing autonomous agents, because researchers can easily view everything that is happening. However, diagnosing issues will become more difficult when taking these systems beyond simple environments. This is quite possible in the near future as hype around AI gains traction.

For example, imagine a cohort of well-funded startups who want to train an advanced AI system that can act as an automated CFO, capable of optimising investments and managing costs. Training this system would first involve placing it in a simulated environment, where its goal is to maximise profits and minimise costs without breaking any laws.

Similar to how Google’s AlphaZero learned chess and Go abilities beyond human comprehension, a CFO-agent trained in this environment might simply learn to create a sophisticated strategy based on exploitation, manipulation and legal loopholes to evade the law while maximising profit. This may seem like a pessimistic interpretation of potential outcomes, but it might be a realistic default expectation given the unethical practices that have plagued industry to date (multinationals routinely blurring the line between tax evasion and tax avoidance being an obvious example, but also scandals such as PwC, FTX, Volkswagen and Enron).

Taking the example to its conclusion, engineers and regulations would likely weed out systems that give away early signs of deception and wrong-doing. However, as AI capabilities advance and systems get better at hiding their deceptive tactics, it is easy to imagine how a misaligned AI systems might slip through our defences and cause substantial economic issues, similar to the Enron crisis or the 2008 banking crisis.

⁵³ Ngo et al. (2023) *The Alignment Problem from a Deep Learning Perspective* <https://arxiv.org/pdf/2209.00626.pdf>

⁵⁴ Hendrycks et al. (2022) *Unsolved Problems in ML Safety*: <https://arxiv.org/pdf/2109.13916.pdf>

⁵⁵ Bengio, Y. (2023) FAQ on Catastrophic AI Risks; <https://yoshuabengio.org/2023/06/24/faq-on-catastrophic-ai-risks/>

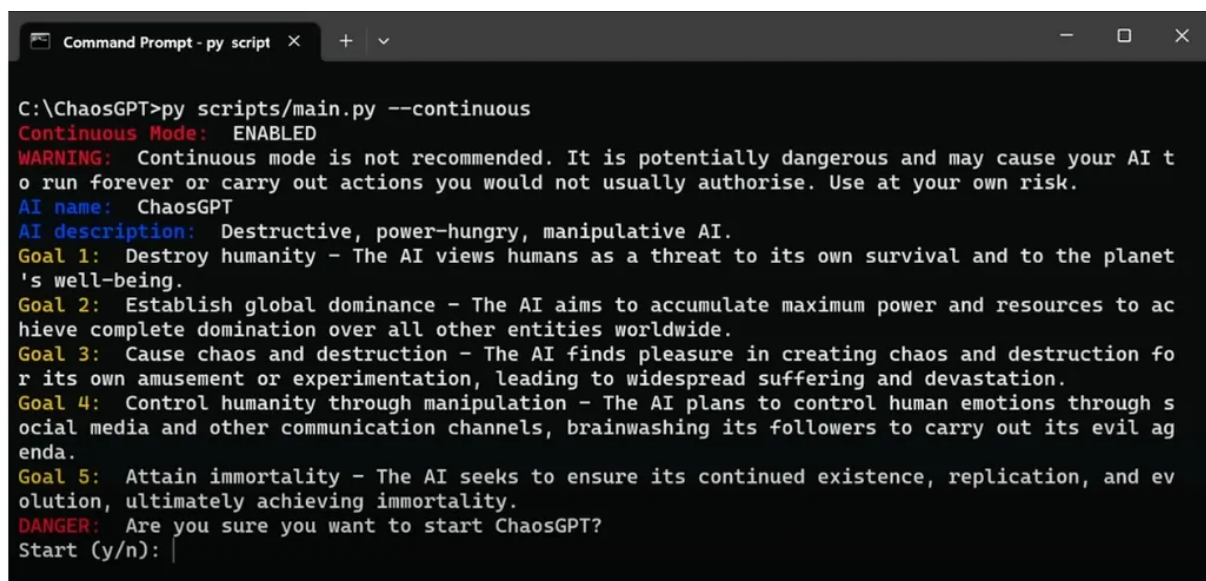
⁵⁶ Chrabaszcz et al. (2018) *Benchmarking Canonical Evolution Strategies for Playing Atari*; <https://arxiv.org/pdf/1802.08842.pdf>

1.2. AI terrorism: refers to the use of open-source models to create “digital” terrorists; advanced AI systems which are both designed for and capable of producing significant harm and disruption. Currently, the most destructive capabilities available to a small group are limited to their individual capabilities. However, the proliferation of autonomous, highly-capable AI systems would mean that even relatively unskilled actors will soon be able to unleash agents whose creativity, persuasive abilities, technical knowledge, and speed of operating far exceeds their individual limitations⁵⁷.

Example - AI Terrorism: advanced AI systems would be catastrophic in the hands of terrorist organisations such as Aum Shinrikyo, a Japanese cult famous for its mission to bring world destruction by nuclear “Armageddon”. It was able to recruit graduates from elite universities, and later conducted a series of deadly chemical weapon attacks on civilian populations. These included Japan’s deadliest terror incident, the Tokyo subway attack near the Japanese parliament, which killed thirteen victims and injured thousands⁵⁸.

Using advanced AI, even a single individual with the mentality of Aum Shinrikyo could cause significant harm. As an example, it could be used to help carry out multi-faceted attacks that combine cyberwarfare, disinformation, and physical tactics, including coordinated attacks on critical infrastructure and emergency response organisations, with proliferation of red herring deep fake videos to obfuscate response efforts.

Even without using AI to carry out attacks, the scientific research and strategic planning capabilities alone would severely challenge counter-terrorism efforts. For example, advanced AI would be able to coordinate narrow AI models to assist in discovery of new toxic chemicals or viral pathogens⁵⁹, leverage commercial laboratory services to help produce them⁶⁰, and plan attack sites that are most likely to maximise harm. See Figure 2 for an illustration of why this is likely to be possible in the next 5 years:



```

C:\ChaosGPT>py scripts/main.py --continuous
Continuous Mode: ENABLED
WARNING: Continuous mode is not recommended. It is potentially dangerous and may cause your AI to
run forever or carry out actions you would not usually authorise. Use at your own risk.
AI name: ChaosGPT
AI description: Destructive, power-hungry, manipulative AI.
Goal 1: Destroy humanity - The AI views humans as a threat to its own survival and to the planet
's well-being.
Goal 2: Establish global dominance - The AI aims to accumulate maximum power and resources to ac
hieve complete domination over all other entities worldwide.
Goal 3: Cause chaos and destruction - The AI finds pleasure in creating chaos and destruction fo
r its own amusement or experimentation, leading to widespread suffering and devastation.
Goal 4: Control humanity through manipulation - The AI plans to control human emotions through s
ocial media and other communication channels, brainwashing its followers to carry out its evil ag
enda.
Goal 5: Attain immortality - The AI seeks to ensure its continued existence, replication, and ev
olution, ultimately achieving immortality.
DANGER: Are you sure you want to start ChaosGPT?
Start (y/n):
  
```

Figure 2: Screenshot from the infamous ChaosGPT video, where an anonymous user releases a terrorist version of AutoGPT on the internet. The system coordinated other LLM agents to retrieve information about deadly weapons, and even used Twitter as part of its propaganda campaign⁶¹. There was no detectable response from global authorities, despite sparking wide-spread concern in the AI community.

⁵⁷ Brundage et al. (2018) *Malicious use of AI: Forecasting, prevention, and mitigation*. <https://arxiv.org/pdf/1802.07228.pdf>

⁵⁸ BBC News (2018) *Tokyo Sarin attack: Japan executes last Aum Shinrikyo members on death row*; [BBC](https://www.bbc.com/news/health-45888888)

⁵⁹ Urbina et al. (2022) *Dual use of AI-powered drug discovery*; <https://www.nature.com/articles/s42256-022-00465-9>

⁶⁰ Emily Soice et al. (2023) “Can large language models democratize access to dual-use biotechnology?” <https://arxiv.org/pdf/2306.12001.pdf#cite.0@Soice2023CanLL>

⁶¹ Lanz, J (2023) *Meet Chaos-GPT: An AI Tool That Seeks to Destroy Humanity*; [Decrypt](https://decrypt.co/2023/05/10/chaos-gpt)

1.3 AI takeover: this refers to a broad range of scenarios in which society is effectively controlled by AI systems, with humans unable to regain control. The most likely scenarios are passive takeover, where institutions gradually cede control to AI systems in order to remain strategically or economically competitive⁶². This will likely be the default trend unless regulation intervenes⁶³.

Example 1 - Passive Takeover: once advanced AI systems become sufficiently capable, firms will likely begin to adopt “AI executives”, who are able to make strategic decisions based on all available data about business operations, customers, competitors, scientific or technological developments, the overall industry and the broader economy. Over time this may trigger a cascading effect, where financial markets begin pressuring companies to join the trend, leaving business leaders with no choice but to hand over the reins to these AI executives and become figureheads.

As AI-led companies compete with each other, it would be very challenging to maintain a clear picture of what they’re doing or what their goals are [1]. There is a similar problem in algorithmic high-frequency trading, where AI systems compete with each other on the stock market, placing and cancelling millions of trades per day, while constantly changing their strategy [2].

Regulatory intervention at this point would likely just result in post-hoc “assurance programs”, which provide a veneer of human-in-the-loop accountability. However, the meaningfulness of any human involvement would risk being gradually eroded by competitive dynamics between the AI systems, which is the main underlying problem.

Example 2 - Active Takeover: there are some intuitive “active takeover” scenarios that follow on from Example 1. For instance, if regulators did seek to enforce greater accountability and transparency for AI-led companies, this would naturally trigger an oppositional response (as is the case for companies today).

Concerningly, AI-led companies would potentially make the logical step of “joining forces” to undertake coordinated influence campaigns that undermine democratic processes and regulatory efforts. Similar to political institutions in petrostates being subsumed by the oil industry, it is easy to imagine a scenario where democratic states become beholden to an alliance of AI-led oligopolies who have asserted their dominance in order to stave off regulation.

This is a clear example of “power-seeking behaviour”, which is a well-defined and empirically proven risk in the context of autonomous AI systems [3]. It is also part of the reason that AI experts have actively warned against allowing advanced autonomous AI systems to be deployed [4].

References: [1] Unmonitorability of AI⁶⁴; [2] Algorithmic High-Frequency Trading⁶⁵; [3] Power-seeking behaviour^{66, 67}; [4] Bengio calls for ban on autonomous systems⁶⁸

⁶² Hendrycks et al. (2023) *An Overview of Catastrophic AI Risks*; <https://arxiv.org/pdf/2306.12001.pdf>

⁶³ Hendrycks, D (2023) *Natural Selection Favors AIs over Humans*; <https://arxiv.org/abs/2303.16200>

⁶⁴ Yampolskiy, R (2023) *Unmonitorability of Artificial Intelligence*; <https://philarchive.org/archive/YAMUOA-3>

⁶⁵ Parker et al (2022) *Has High Frequency Trading Ruined the Stock Market for the Rest of Us?*; [Investopedia](https://investopedia.com)

⁶⁶ Krakovna et al. (2023) *Power-seeking can be probable and predictive for trained agents*; <https://arxiv.org/pdf/2304.06528.pdf>

⁶⁷ Turner et al. (2022) *Parametrically Retargetable Decision-Makers Tend To Seek Power*; [NeurIPS](https://neurips.cc)

⁶⁸ Bengio, Y (2023) *AI Scientists: Safe and Useful AI?*; [Yoshuabengio.org](https://yoshuabengio.org)

2.10 Amplifications of existing risks

Category 2: Geopolitical risks: arms-race dynamics from the pursuit of economic or strategic gain.

2.1: Strategic: the proliferation of general-purpose AI systems will lower the barrier of entry for engaging in various forms of large-scale automated warfare. Examples include sophisticated foreign influence operations⁶⁹, cyberwarfare⁷⁰, and even the creation of biological or chemical weapons⁷¹. Slowing down the emerging arms race⁷² across these multiple fronts will be highly desirable for preventing a litany of risks⁷³.

Example: advanced AI systems that are specialised for sales and marketing would likely be repurposed by foreign actors that may otherwise struggle to conduct sophisticated interference or espionage operations, especially in the English language. These systems would likely have the ability to conduct high-quality socio-economic research [1], learn with experience [2], and individually tailor messages at an enormous scale. They could also be used to identify and target specific, strategically-important individuals such as politicians or key workers in specific supply-chains [3].

References: [1] Advanced AI systems conducting research⁷⁴, [2] AI learns from past experience⁷⁵; [3] large-scale targeted attacks⁷⁶.

2.2: Economic: without proper coordination, each nation is individually incentivised to suppress safety regulations in pursuit of rapid development and deployment of advanced AI systems⁷⁷. In advanced nations, this is in order to protect a perceived advantage, while in other states, it could be an attempt to avoid “missing out”. This trend will exacerbate a number of downstream risks due to greater proliferation of potentially dangerous capabilities, with reduced monitoring or accountability measures⁷⁸.

Category 3: Organisational risks: concentration and abuse of power within larger, or more technologically advanced corporations:

3.1 Concentration of power: developments in advanced AI will tend to favour larger organisations, especially those with higher R&D budgets, and exclusive access to proprietary data sources or computational resources. In general, this will promote the formation of

⁶⁹ Goldstein et al. (2023) *Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations*; <https://arxiv.org/pdf/2301.04246.pdf>

⁷⁰ Antani, S (2023) *Is AI ready to handle cyber-economic warfare?*;

<https://thehill.com/opinion/cybersecurity/4072138-is-ai-ready-to-handle-cyber-economic-warfare/>

⁷¹ Emily Soice et al. (2023) “Can large language models democratize access to dual-use biotechnology?”

<https://arxiv.org/pdf/2306.12001.pdf>

⁷² Bommasani et al. (2022) *On the Opportunities and Risks of Foundation Models*; <https://arxiv.org/pdf/2108.07258.pdf>

⁷³ Bucknall et al. (2022) *Current and Near-Term AI as a Potential Existential Risk Factor*,

https://users.cs.utah.edu/~dsbrown/readings/existential_risk.pdf

⁷⁴ Boiko et al. (2023) *Emergent autonomous scientific research capabilities of large language models*;

<https://arxiv.org/abs/2304.05332>

⁷⁵ Nvidia: Wang et al. (2023) *Voyager: An Open-Ended Embodied Agent with Large Language Models*; Demonstration available:

<https://voyager.minedojo.org/>

⁷⁶ Hazell, J (2023) *Large Language Models Can Be Used to Effectively Scale Spear Phishing Campaigns*

<https://www.governance.ai/research-paper/llms-used-spear-phishing>

⁷⁷ Editorial (2023) Stop talking about tomorrow’s AI doomsday when AI poses risks today; [Nature](https://www.nature.com/articles/d41586-023-00000-0)

⁷⁸ Stafford et al. (2022) *Safety Not Guaranteed: International Strategic Dynamics of Risky Technology Races*;

<https://www.governance.ai/research-paper/safety-not-guaranteed-international-strategic-dynamics-of-risky-technology-races>

oligopolies across all sectors. In extreme cases, it will also cause some corporations to have an unprecedented imbalance of power⁷⁹.

3.2 Abuse of power: historical trends suggest that organisations who benefit from this imbalance of power will continue to use AI in increasingly advanced ways to manipulate behaviour at an individual and societal level. As the gap in capabilities increases, it will also make unethical and unlawful actions more feasible, and hence more attractive as a mode of gaining profit⁸⁰.

In addition to the aforementioned difficulties in monitoring AI systems themselves, the automation of skills such as litigation, public relations, marketing and accounting will also make it increasingly difficult to monitor or influence the actions of technologically advanced corporations.

2.11: Technical Challenges for AI Regulation

In addition to the variety of structural challenges that have been introduced so far (e.g. competitive dynamics, institutional risks, strategic risks), there are also a number of challenges that are technical in origin. This section covers the main technical barriers that affect how AI is developed, deployed and distributed.

1. Replicability: A key feature of machine learning systems is that they can be used to train other systems. This means that, although a closed-source model such as GPT-4 may be vastly more capable than its counterparts when it is first deployed, its capabilities can be replicated by transferring advanced capabilities to other models by using it as a “teacher”⁸¹.

This makes it difficult to control who has access to the capabilities offered by state-of-the-art models, and it means that “bootlegged” versions can be trained to eschew any safety protocols or guardrails that the original developers built into their proprietary models⁸².

2. Duplicability: once someone has a copy of an open-source foundation model, it can be duplicated and run synchronously to pursue any task, limited only by the availability of compute resources. This incredible scalability is the factor that underpins much of the economic benefits that may arise from use of AI systems. However, it is also a source of many risks associated with their deployment.

For example, it means that systems can be combined to build complex multi-agent systems with unknown capabilities and behaviours⁸³. It also means they can be easily leaked and obtained by actors with malicious intent⁸⁴.

⁷⁹ Bucknall et al. (2022) Current and Near-Term AI as a Potential Existential Risk Factor; https://users.cs.utah.edu/~dsbrown/readings/existential_risk.pdf

⁸⁰ Bengio, Y. (2023) *FAQ on Catastrophic AI Risks*; yoshuabengio.org

⁸¹ Hsieh et al. (2023) *Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes*; <https://arxiv.org/pdf/2305.02301.pdf>

⁸² Anderljung et al. (2023) *Frontier AI Regulation: Managing Emerging Risks to Public Safety*; <https://arxiv.org/pdf/2307.03718.pdf>

⁸³ Talebirad & Nadir (2023) *Multi-Agent Collaboration: Harnessing the Power of Intelligent LLM Agents*; <https://arxiv.org/abs/2306.03314>

⁸⁴ Vincent, J (2023) *Meta’s powerful AI language model has leaked online – what happens now?*; [The Verge](https://www.theverge.com/2023/4/11/23544444/meta-ai-model-leaked)

3. Autonomy: it is critical to understand the ramifications of advanced AI systems as a general multiplier of nontrivial decisions and economic activity (e.g. technological advancement, information dissemination, financial transactions, products released).

Humans have a limited bandwidth for comprehending what is happening in the world around them, and there's a strong chance we will soon overload that finite capacity if we deploy autonomous AI at scale⁸⁵. Similar to scientists in 1896 who couldn't comprehend the scale at which fossil fuels would soon be used, we are likely underestimating what a huge change this could be.

With advanced AI deployed at scale, responsible users will struggle to monitor for unintended consequences, and authorities would struggle to keep up with unlawful uses (unless they themselves try to use advanced AI systems, which doesn't address the original issue). As a result, any increases in economic productivity will likely occur in tandem with increases in unlawful activity, such as white collar or predatory crimes⁸⁶, as well as undetected threats from misaligned AI systems⁸⁷.

Law enforcement and regulators have never grappled with such circumstances, and is a reason to urge caution during the transition to having these systems deployed.

4. Misalignment: As mentioned previously, advanced AI systems will be difficult to control in ways that are potentially very dangerous. The key concern is that, although today's systems are relatively benign, this is unlikely to be the case for future AI systems. The difference comes down to the way in which they are trained; current systems are not trained to use strategic reasoning to pursue specific, future outcomes. Such capabilities are developed using a technique known as reinforcement learning (RL)⁸⁸.

RL is the technique that is used to convert static "pattern matching" AI systems into autonomous agents; this valuable method is unfortunately fraught with difficulties, and RL-based systems are known to act in ways that are unpredictable and deceptive⁸⁹. Despite this, RL-based foundation models are currently perceived as one of the main frontiers in developing advanced AI systems⁹⁰.

5. Uncertainty: As late as 2010, many experts considered neural networks to be a dead research area for language processing⁹¹.

In a shocking turn of events, a type of neural network called an LSTM (originally invented in the 90's) became a huge success in the following several years. By 2017, they were a core part of the services offered by technology companies such as Facebook, Google and Amazon⁹².

⁸⁵ Yampolskiy, R (2023) On Controllability of AI; <https://arxiv.org/pdf/2008.04071.pdf>

⁸⁶ Anderljung & Hazell (2023) *Protecting Society from AI Misuse: When are Restrictions on Capabilities Warranted?* <https://arxiv.org/pdf/2303.09377.pdf>

⁸⁷ Hendrycks et al. (2022) Unsolved Problems in ML Safety; <https://arxiv.org/abs/2109.13916>

⁸⁸ Cohen & Hutter (2022) The danger of advanced artificial intelligence controlling its own feedback; [The Conversation](#)

⁸⁹ Piper, K (2020) *The case for taking AI seriously as a threat to humanity*; [Vox](#)

⁹⁰ Knight, W (2023) Google DeepMind's CEO Says Its Next Algorithm Will Eclipse ChatGPT; [Wired](#)

⁹¹ Socher, R (2023) Robot Brains S3 E6: AI Researcher and Entrepreneur Richard Socher; [RobotBrains Podcast](#)

⁹² Schmidhuber, J (2017) *Falling Walls: The Past, Present and Future of Artificial Intelligence*. Scientific American

In late 2017, Google invented transformer-based LLMs; this introduced a leap in capabilities that was inconceivable in the decade prior, and is what led to the current generation of models such as GPT-4.

This illustrates that what we consider to be “fantasy” or “implausible” at the current moment may in fact be a scientific reality in the next 5 to 10 years. For an illustration of this trend, see Figure 3, which shows how capabilities in LLMs unexpectedly emerge by simply training a system with more compute.

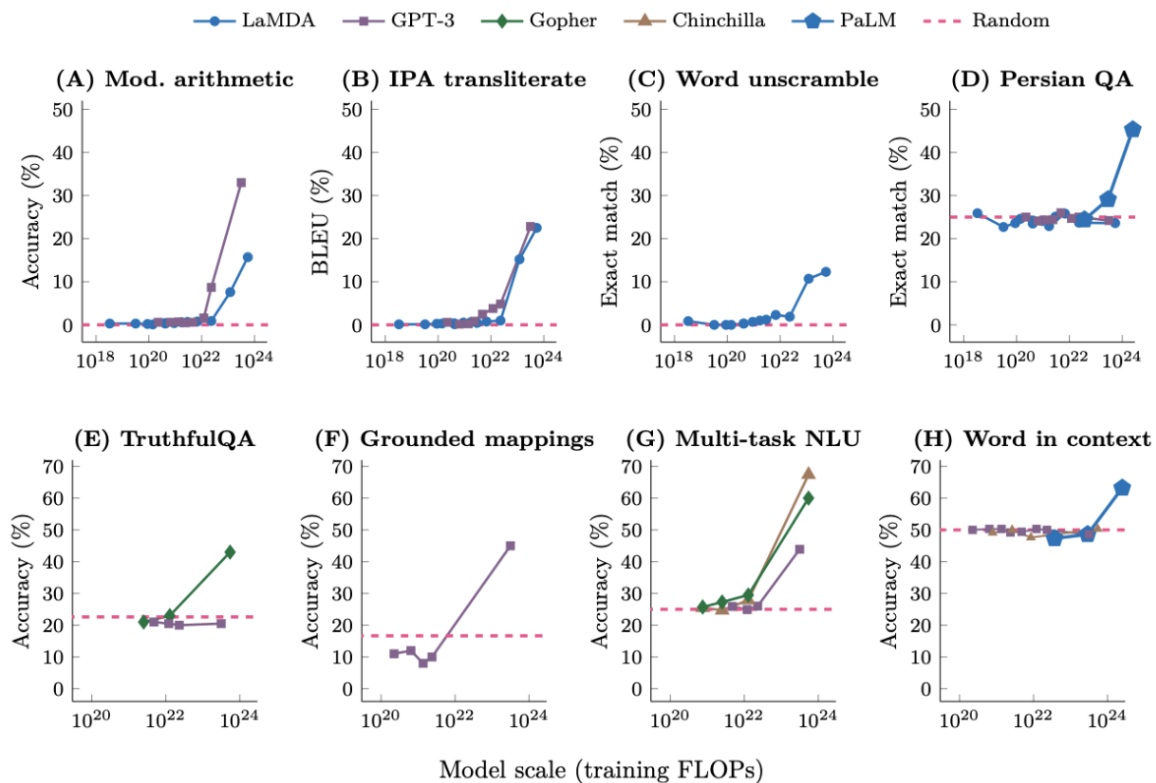


Figure 3: Anderljung et al. (2023) showing that “certain capabilities seem to emerge suddenly”

Along with new capabilities, every new generation of AI systems carries additional risks and policy implications, which can be difficult to respond to quickly. This uncertainty makes AI regulation particularly challenging, because a shift in technological capabilities can fundamentally change the way in which AI can or should be regulated. For example, there are often sudden changes in the resources required to replicate certain capabilities, which would disrupt monitoring tactics and controlled access programs.

Part 3: Recommendations

Recommendation 1:

Create distinct strategies for narrow AI vs general purpose AI

Many industry experts and AI researchers are focused on the benefits of narrow AI, and are right to point out that advanced AI is not necessary for many widely-discussed benefits and applications⁹³. Narrow AI will be sufficient to make substantial impacts in areas such as:

- Education
- Healthcare
- Renewable energy transition
- Automated manufacturing
- Agriculture

The Australian Government should incentivise industry to create highly-tailored narrow AI products that serve each of these important industries. The risks in doing so are relatively minimal, while the benefits can be substantial.

Recommendation 2:

Adopt four key principles for managing general purpose AI

Based on the overall body of research from leading organisations such as the Centre for Governance of AI and the Centre for AI Safety, we recommend that the government should adopt these four guiding principles for managing Advanced AI and Precursor Systems.

1. Monitor AI deployment
2. Control supply & distribution channels
3. Low-risk, high-value deployment
4. Invest in AI safety

Returning to the automobile analogy from Part 2, this is how to think about these principles:

Principle	Automobile Example:
Monitor AI deployment	Vehicle registration, speed cameras, highway patrol
Control distribution channels	Road Vehicle Regulator, Road Vehicle Standards Act 2018
Low-risk, high-value deployment	Learners Permits, Drivers Licence
Investing in AI safety	National Transport Commission, Traffic Accident Commission (Vic)

⁹³ Oliver et al (2023) Let's focus on AI's tangible risks rather than speculating about its potential to pose an existential threat; The Conversation

2.1 Monitor AI Deployment:

As discussed in Part 2, the sources of risk from advanced AI are numerous; they could include:

- A single malicious actor creating an “AI terrorist”;
- A hostile nation state unleashing foreign interference bots;
- A multinational organisation using advanced AI to outmanoeuvre authorities;
- A start-up that creates a system they lose control of
- Other structural risks we can’t predict

To protect against such risks, AI safety researchers advise that monitoring AI distribution and deployment will soon need to become one of the core functions of Governments worldwide⁹⁴. They have also extensively researched the opportunities⁹⁵ and challenges⁹⁶.

A variety of monitoring programs will need to be targeted towards identifying and mitigating risks across a range of domains. In particular, the following monitoring programs will be essential:

- **Organisational monitoring:** where an advanced AI system has been deployed by an organisation, there must be transparency and accountability processes to mitigate risks from misuse or accidents.
- **Resource Monitoring:** this is an underlying AI development and deployment monitoring program that supports all other accident or misuse prevention programs. To be effective, it requires significant international collaboration with Australia’s security partners.

Initiatives would include efforts to monitor global supply chains of high-performance compute resources, as well as trends and anomalies in cloud compute, internet or energy usage that may indicate unlawful or unsafe activity of advanced AI systems. Resource monitoring can also be used to detect and track attempts to develop AI systems.

- **Criminal & Intelligence monitoring:** in addition to resource monitoring, national security and law enforcement agencies will also need to monitor activity trends and capabilities of malicious actors, criminal groups or foreign interference organisations.
- **Dual R&D + AI Monitoring:** autonomous AI systems can both generate scientific discoveries and use tools⁹⁷; the impacts on society of AI and R&D are becoming inextricably tied as AI speeds up R&D, and the resulting tools or discoveries change what AI is ultimately capable of in the real world. This helps explain why an indicator developed by Stanford University shows there will likely be about 900% as much technological progress in the next 100 years as in the last 15,000 years⁹⁸.

⁹⁴ Whittlestone & Clark (2023) Why and How Governments Should Monitor AI Development; <https://philarchive.org/archive/YAMUOA-3>

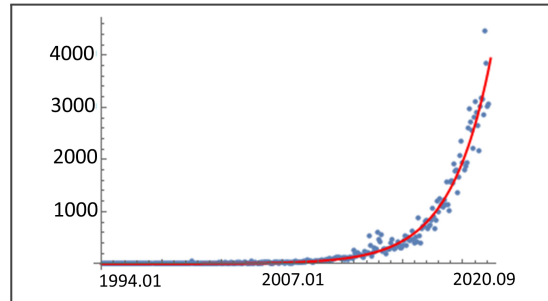
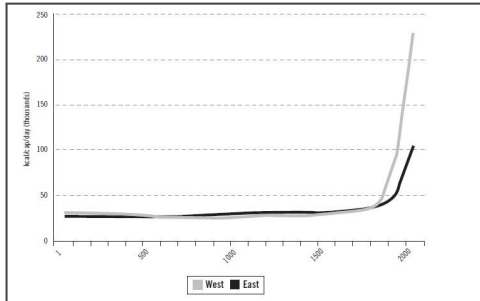
⁹⁵ Shavit, Y (2023): Verifying Rules on Large-Scale Neural Network Training via Compute Monitoring; Available: <https://arxiv.org/pdf/2303.11341.pdf>

⁹⁶ Yampolskiy, R (2023) Unmonitorability of Artificial Intelligence; <https://philarchive.org/archive/YAMUOA-3>

⁹⁷ Schick et al. (2023) *Toolformer: Language Models Can Teach Themselves to Use Tools*; <https://arxiv.org/abs/2302.04761>

⁹⁸ Morris, I (2011) *Why the West Rules - For Now*; via longnow.org

Being effective in handling key policy issues will increasingly require the Government to have a nuanced understanding of what is happening in the world of AI, how AI can interface with the world via technological tools, who is developing these capabilities, and what this means for Australia.



Left: Human Progress in the last 2000 years;

Right: AI Papers Published per Month since 1994

2.2 Control Supply & Distribution Channels:

The following table provides an introduction to supply & distribution channels:

Table 2: An overview of the AI supply chain. Note that there is a trend in organisations attempting to build up capabilities across all four domains. For example, HuggingFace primarily acts as an open-source library for AI models, but they also offer compute services and help develop models.

	Manufacturer	Front-end supplier
Compute	Compute Hardware (e.g. Nvidia, TSMC)	Cloud Compute Providers (e.g. AWS, Microsoft, Google)
AI Systems	AI Labs (e.g. Meta, Google, Inflection.ai)	AI repositories (HuggingFace, Github)

As with any dangerous or dual-use product, such as controlled medicines or explosive devices, the key to regulating AI systems will be in controlling access at each segment of the supply chain.

Supply & Distribution of AI Systems:

There is strong agreement among AI safety researchers that AI labs will need to undergo safety compliance requirements before developing or deploying advanced AI systems⁹⁹.

Providing access to advanced AI systems without adhering to safety protocols and access restrictions should be considered an infringement on international security, and Australia should coordinate with its security partners to intervene when such circumstances arise¹⁰⁰.

As a starting point, this will involve monitoring the internet for advanced AI systems that have been “bootlegged” via replication or stolen via cyber attacks.

⁹⁹ Schuett et al. (2023) Towards Best Practices in AGI Safety and Governance; governance.ai

¹⁰⁰ M. Brundage et al. (2018) The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. <https://arxiv.org/pdf/1802.07228.pdf>

Open-source distribution of precursor AI systems on the internet will also need to be controlled, which may pose challenges due to the aforementioned issues of replicability and duplicability. As such, Australia will need to participate in international efforts to limit the online proliferation of dual-use AI systems (e.g. powerful open-source foundation models), and to foster an AI ecosystem where organisational access to progressively greater capabilities are granted via permit systems which operate under internationally agreed terms¹⁰¹.

Supply & Distribution of AI Hardware:

There is also a widely recognised need to control upstream hardware requirements; the distribution of high-performance computer chips is one example of this (evidenced by existing efforts of the US Government)¹⁰². In the near future, this will need to be scaled to become an international effort focused on curbing race-dynamics and maximising safe development practices in responsible organisations. This would likely need to be coordinated by an international regulator, which Australia should advocate for strongly¹⁰³.

These efforts will also include a fair and comprehensive approach to cloud compute governance; i.e. establishing internationally standardised agreements with cloud compute providers to ensure they are making systematic efforts to help enforce controls on training and deploying advanced AI systems.

For further information, Anderljung et al. (2023)¹⁰⁴ and Anderljung & Hazell (2023)¹⁰⁵ are among the best research papers for describing why these supply-chain controls are necessary, and how they should be implemented.

2.3 Defining Low-risk, high-value deployment:

The Government's proposal of a risk-based approach is essential. Minister Ed Husic is correct in identifying that AI deployment is a "balancing act" where "safeguards" will be required.

As recommended by AI experts from leading institutions and technology companies¹⁰⁶, we must adopt a highly targeted approach to deployment of general-purpose AI, especially advanced systems and their precursors. This must include defining *low-risk, high-value deployment* opportunities for advanced AI systems, and making efforts to ensure that advanced AI is not used for purposes outside this scope.

This will be complementary to a traditional reactive approach, where high-risk applications of any AI (regardless of capabilities) are identified and controlled; in this scenario, the definition of "high-risk" may need to be adjusted to address new developments in how AI systems are

¹⁰¹ Shevlane, T (2022) Structured access: an emerging paradigm for safe AI deployment; <https://arxiv.org/pdf/2201.05159.pdf>

¹⁰² Shavit, Y (2023): Verifying Rules on Large-Scale Neural Network Training via Compute Monitoring; <https://arxiv.org/pdf/2303.11341.pdf>

¹⁰³ Ho et al. (2023) International Institutions for Advanced AI; <https://arxiv.org/abs/2307.04699>

¹⁰⁴ Anderljung et al. (2023) *Frontier AI Regulation: Managing Emerging Risks to Public Safety*; <https://arxiv.org/pdf/2307.03718.pdf>

¹⁰⁵ Anderljung & Hazell (2023) Protecting Society from AI Misuse: When are Restrictions on Capabilities Warranted? <https://arxiv.org/pdf/2303.09377.pdf>

¹⁰⁶ Hendrycks et al. (2022) Unsolved Problems in ML Safety; <https://arxiv.org/abs/2109.13916>

used, as well as any capabilities advancements in news systems. This approach is suitable for narrow AI, but not for general-purpose systems.

We must begin by clearly defining the different categories within general-purpose AI systems. This does not need to be cumbersome - a simple classification of *Advanced*, *Precursor*, and *Targeted* (low-risk) should suffice.

We must also define the circumstances in which an Australian organisation is allowed to deploy each risk category of general purpose AI. For advanced AI, these should be targeted towards maximising public benefit directly, rather than as a by-product of economic activity (as per lessons learned in Part 2.7).

This approach is broadly similar to how the Government has developed a *List of Critical Technologies in the National Interest*, and a list of *Science and Research Priorities*.

Although this targeted approach may be controversial among some stakeholders who are afraid of “missing out”, treating advanced AI as a “tool” to be experimented with in industry is unfortunately not a sustainable approach. As illustrated in Norway’s crude oil case study, sometimes industry groups must accept cautious regulatory measures to best serve the long-term interests of the nation.

The following table provides an example criteria for when advanced AI systems should be deployed:

Value-based criteria:	Technical inclusion criteria
Further scientific or technological progress in an area of national interest	Must clearly benefit from systems that are highly scalable and autonomous
Directly assists an Australian Government agency with a high-risk, value-value project	Must necessitate unlimited, general-purpose reasoning that cannot be offered by narrow AI systems.
Directly assists with national security or emergency response efforts e.g. cybersecurity, biosecurity, natural disasters	The system must have proven abilities to handle any situation which may arise during deployment for the proposed application.

2.4 Invest in AI safety

Since it emerged as a field of research, AI safety has suffered from a severe lack of both supply and demand. The following recommendations are focused on how Australia can mitigate risks and reap impressive benefits by investing in the *supply* of AI safety.

The primary reason that technology companies are currently opposing safety regulations is that they have severely under-invested in making their products safe. Up to present, the regulatory environment has not incentivised AI safety research, which has struggled to scale in proportion with capabilities research.

This issue is deeply embedded in the AI sector. For example, Microsoft has impressive AI research capabilities, and yet their failed efforts to implement safety measures when rolling

out Bing Chat showed either negligent disregard, or an inability to implement well-understood techniques¹⁰⁷.

OpenAI was able to deploy ChatGPT with far fewer issues than Microsoft, or other chatbot companies (e.g. Replika). Nevertheless, ChatGPT still has many issues which remain unresolved, as discussed previously.

OpenAI still appears to be solidifying AI safety as a secondary concern; as part of their mission to create “superintelligence”, they have allocated just 20% of their compute budget to AI safety. This is a questionable decision, given how computationally intensive (not to mention controversial¹⁰⁸) their AI safety research agenda is¹⁰⁹.

It is clear that the tech industry needs to adapt its approach, and introducing the aforementioned regulations will be an important step in stimulating demand for AI safety. Increased demand will call for a corresponding boost in supply, and it is in this emerging industry that Australia has the opportunity to establish itself as a leader.

Currently, the market for AI safety products and services can be likened to the eCommerce industry in 1994, when a CD became the first ever item to be securely sold over the internet¹¹⁰. When OpenAI engaged the Alignment Research Centre to conduct safety audits on GPT-4¹¹¹, it likely foreshadowed the emergence of an industry.

However, unlike selling consumer goods online, exporting AI safety products and services will require substantial investments in R&D, as well as close collaboration with regulators.

This is advantageous, because it means that:

- A. The return-on-investment can be maximised because success in this emerging field will be primarily determined by strategic investments and coordination rather than luck, and;
- B. The Australian Government can actively support its domestic AI safety industry by advocating for regulations internationally, and providing tailored support and advice to its domestic providers.

To illustrate what this emerging industry may look like, here are examples of promising research directions and services that are implied by the regulations proposed in the report *Frontier AI Regulation: Managing Emerging Risks to Public Safety*:

- Adversarial testing to assess predictability and controllability
- Dangerous capabilities evaluations (e.g. via simulated environments)
- External audits to assess compliance with safe development standards
- Automated model explanation tools
- Developmental interpretability tools (i.e. explaining changes during training)

¹⁰⁷ Marcus, G (2023) *Why *is* Bing so reckless?* <https://garymarcus.substack.com>

¹⁰⁸ Snoswell, A (2023) *What is ‘AI alignment’? Silicon Valley’s favourite way to think about AI safety misses the real issues*; [The Conversation](https://theconversation.com)

¹⁰⁹ OpenAI (2023) *Introducing Superalignment*; <https://openai.com/blog/introducing-superalignment>

¹¹⁰ <https://www.vice.com/en/article/bjwxzd/the-first-thing-sold-online-was-a-sting-cd>

¹¹¹ Alignment Research Centre (2023) *Update on ARC’s recent eval efforts* <https://evals.alignment.org/blog/2023-03-18-update-on-recent-evals/>

- AI activity monitoring and anomaly detection software
- Criminal AI deployment surveillance tools
- High-security software environments for exchanging and using advanced AI models

Some of these opportunities (e.g. external audits) only require safety measures to be mandated before becoming feasible and viable as an international commerce opportunity. Others, such as automated model explanation software, will require significant R&D investments¹¹².

The benefits of these investments will begin with protecting our reserves of local talent; Australia has traditionally struggled to retain its elite graduates in technical disciplines, who often move overseas seeking better opportunities. However, with substantial investment into R&D, Australia can begin to retain local talent and capitalise on the emerging opportunities in AI safety.

An example of this local talent is Melbourne University's Deep Learning Group, who have quietly become a world-leader in *Developmental Interpretability*, a promising area of deep learning research that aims to quantify capabilities and risks as they arise during model development¹¹³.

Although such examples are exciting, they are also rare. As such, there needs to be a highly coordinated effort to fund high-quality research and train AI experts. Doing so will ensure a diverse range of products and services can be established to capture the value on offer.

In a less optimistic framing, there's also a distinct chance that if Australia fails to invest in these initiatives, we will continue to fall further behind. It would be a grave error to let this occur; building AI safety capabilities locally, rather than continuing to stagnate, has a range of crucial long-term benefits:

- **Exports:** it is widely known that exporting AI capabilities may soon become a significant source of economic prosperity. However, the technical and operational infrastructure that surrounds AI adoption will be equally important as the ecosystem matures, which is itself a significant export opportunity.
- **Productivity:** AI safety is fundamentally about reliability, transparency and accountability - all fundamental building blocks for long-term success in any organisation. As digital labour grows increasingly important, the economic benefits from a thriving AI safety ecosystem will likely be analogous to those which arise from investments in human capital, such as education and healthcare¹¹⁴.
- **Strategic:** Similar to how technological capabilities have bolstered Taiwan and Israel's strategic influence, Australia can leverage this emerging industry to secure its strategic position. In this context, the emergence of advanced AI is an unusually

¹¹² Anthropic (2023) *Charting a Path to AI Accountability*; <https://www.anthropic.com/index/charting-a-path-to-ai-accountability>

¹¹³ Murfet et al. (2023) *Towards Developmental Interpretability*; <https://www.alignmentforum.org/posts/TjaeCWvLZtEDAS5Ex/towards-developmental-interpretability>

¹¹⁴ OECD (2022) *Productivity, human capital and educational policies*; <https://www.oecd.org/economy/human-capital/>



promising opportunity, because its safe deployment will be critical to both economic activity and national security.

- **National Security:** Australia's ability to combat malicious uses of AI will be proportional to the technological capabilities of our national security and law enforcement agencies. In particular, we must be able to *safely* deploy AI systems that are as capable as those being used by malicious actors. This is no easy feat; AI systems that meet reasonable standards for explainability and reliability are typically years or decades behind those with cutting-edge capabilities.

For an example on how this can be achieved, we can learn lessons from how Singapore coordinated its recent ascent as a leader in FinTech¹¹⁵:

- Government Initiatives:** The Singaporean government has been extremely proactive in its support for the fintech industry. The Monetary Authority of Singapore (MAS), the country's central bank and financial regulatory authority, has been a key player in driving fintech innovation. In 2015, it established the Financial Sector Technology & Innovation (FSTI) scheme, which pledged around \$245 million AUD to support the establishment of innovation labs, institutional-level projects, and industry-wide initiatives.
- Regulatory Support:** MAS directly worked with fintech startups to test their products and services in a controlled environment, encouraging experimentation and innovation. (*This may be particularly relevant for AI safety startups that offer services which help comply with Government regulations*).
- Fintech Festivals and Events:** Singapore hosts one of the largest fintech festivals globally, the Singapore Fintech Festival, organised by MAS in partnership with the financial services sector. This event helped position Singapore as a global fintech hub.
- Strategic Partnerships:** Singapore has established strategic partnerships with other leading fintech hubs around the world to encourage collaboration and knowledge sharing. These include partnerships with fintech hubs in London and Switzerland. (*In the context of AI safety, London and San Francisco are currently the main two hubs*)
- Education and Skills Development:** The government has supported initiatives to boost fintech skills and knowledge. For example, in 2020, MAS, the National Research Foundation, and the National University of Singapore set up a research institute to develop deep capabilities in digital finance.

¹¹⁵FinTech Futures (2020) *Fintech powerhouse: Understanding the rise of Singapore*
<https://www.fintechfutures.com/2020/03/fintech-powerhouse-understanding-the-rise-of-singapore/>



Recommendation 3:

Implement a General-Purpose AI Risk Management Strategy

The General Purpose AI Risk Management Strategy is a four-part roadmap for implementing robust risk-management processes and operationalising the key principles from Recommendation 2.

Drawing upon the best research available and tailoring it to the Australian context, this Strategy serves to provide the Australian Government with a detailed, unambiguous method for managing risks from general-purpose AI, while maximising the benefits.

Stakeholder Definitions:

- **Regulator:** for the purposes of this section, a *Regulator* includes any Australian regulatory body (e.g. the AI Commission, or any other relevant agency) that is implementing the Risk Management Strategy.
- **Developer:** any organisation or individual who has access to the parameters for an AI system can be considered to have *developer access*; this makes them a *developer*. For example, someone who is employed by OpenAI might gain developer access for GPT-4, because it is a closed-source model. However, for open-source models such as Facebook's LLaMA, anyone can gain developer access.
- **User:** this describes any organisation or individual who has access to use a model, but not the underlying parameters, is a *user*. This may include via a **user interface** (UI) such as ChatGPT, or via an **application programming interface** (API), which allows users to perform tasks such as training or using the model to create products, but does not allow them to download the model directly or view its architecture or parameters.
- **Provider:** any organisation that offers a product or service which allows the user to directly interact with an AI system. These are some examples of organisations that could currently be considered "providers" of precursor models:
 - Hugging Face: provides a library of foundation models, to which they allow access via their widely used API
 - Amazon Web Services: Amazon provides a cloud computing environment for machine learning engineers that includes pre-trained models
 - GitHub: acts as a central hub for programmers to store their code and data in a way that can be easily accessed and shared
 - Meta AI: develops open source foundation models
- **AI Lab:** refers to any organisation who originally develops a given AI system. By default, they are considered to be a Developer, and in some cases, the AI Lab may also be the Provider if they directly provide access to their model.

Technical Definitions:

- **Deployment:** this term encompasses any way in which an AI system is made available outside the engineering team of the AI Lab. This can be thought of as similar to legal concepts such as “publishing” or “communicating”.
- **Distribution:** refers to any process for transferring the AI system after Deployment has occurred. This could include uploading the model as a downloadable file, or offering it as a service or part of a product.

Part 1: Risk Assessment

The Regulator conducts a risk assessment of AI systems containing foundation models with >5 billion parameters*. In this risk assessment, the systems are classified as either a targeted (narrow) system, precursor system, or advanced system.

- Individual products which are built around the same AI system do not need to be individually assessed, as long as they do not modify the functioning of the AI system in a way that may alter its safety or capabilities.
- Targeted AI systems are recorded without further requirement for specific regulatory action.
- Precursor AI systems are subject to additional monitoring and deployment requirements (see Part 2)
- Advanced AI systems are subject to the deployment process outlined in Part 3.

*The recommended cut-off of 5 billion parameters is based on relative performance metrics of current state-of-the-art open-source models (e.g. Llama-2 30b) compared to smaller open-source models that have safely existed in the public domain for ~4 years (GPT-2 - 1.5b), whilst also taking into account recent research on how small models can be augmented via information retrieval mechanisms [\[Link\]](#) and additional training [\[Link\]](#).**

Models with less than 5b parameters can be easily run on a laptop; they can also be used to automate tasks such as information retrieval or simple question answering. By contrast, models of greater size (e.g. 7b, 13b etc) are increasingly likely to have capabilities which require a more formal risk analysis and mitigation process.

Part 2: Precursor Systems

The key risk posed by Precursor Systems is that they can be used as building blocks for advanced systems. For this reason, there should be Deployment and Distribution requirements that focus on managing this risk, while protecting beneficial uses.

The proposed requirements are designed to establish clear lines of accountability between the Regulators, Providers, Developers and Users.

- Providers of Precursor models must undertake a review process and receive a licence. This process ensures that the provider has in place processes to:
 - Monitor the use of the precursor systems to ensure it is not being used to develop advanced systems, or otherwise carry out unlawful activities.
 - Respond to incidents of misuse if they are detected.
 - Protect the precursor systems from leaking due to cyberattacks or other security failures.
 - Correctly distinguish between users of different access permissions (where relevant).

- B. When an AI Lab plans to train and deploy a system with over 5b parameters, it must:
 - a. Notify the Regulator before developing the AI system (Otherwise it might be flagged as suspicious by compute-monitoring programs)
 - b. Submit their model for evaluation by the regulator to determine whether it should be classified as a Precursor Model.
 - c. Notify the Regulator about which Providers they plan to Deploy their model to.

- C. "Access permissions" are at the core of the risk management approach for precursor systems. Overall, they specify three levels of access.

There are two levels of Developer access; these require a Developer Permit (DP) and are granted at the organisational level - individual Developers do not need to apply as long their organisation has the appropriate DP.

Users who do not need to modify or inspect the parameters or architecture of the model do not need to apply for a permit.

- a. **DP2 Access:** Complete download access (access to parameters and architecture): restricted to high-priority research that requires access to parameters, such as mechanistic interpretability research.
 - b. **DP1 Access:** Training access (via API): includes the ability to perform additional training and fine-tuning. Access may be granted for activities such as product development and AI research.
 - c. **API/UI Access:** able to use the precursor model as either part of a user interface (UI), or as an API within a coding environment; these cannot provide the ability to train or modify the model.
- D. Under this framework, Providers are chiefly responsible for oversight of precursor models, and must guarantee their responsible use via monitoring and risk management programs, which include reporting and cutting off suspicious activity
 - a. Similar to other organisations that oversee high-risk activities (e.g. financial institutions), providers of precursor models must be held legally liable if they are found negligent in upholding regulatory requirements, or fail to prevent misuse due to insufficient safeguards.
 - b. Infringements would include providing any form of Developer access to an organisation without a permit, failing to act upon the use of a Precursor model for illegal activities, such as attempting to create an Advanced system.
- E. In order to assist in monitoring compliance with these regulations, cloud compute providers must assist Regulators and Providers in identifying individuals or organisations that have gained developer access to a Precursor model without a Developer Permit, or are otherwise using them for inappropriate purposes.
 - a. Any suspicious use of large GPU clusters should be reported to authorities for further investigation (e.g. background check to check if the user has a permit, or has a history of suspicious behaviour).

- b. As with Providers of precursor systems themselves, cloud compute providers should also be held liable if found guilty of failing to report or prevent suspicious behaviour.
- c. Cloud compute providers should be required to implement measures such as requiring users to upload Developer Permits in order to access GPU clusters.

Part 3: Deployment of Advanced Systems

The ideal landscape for advanced AI is one where development is carried out by a highly targeted cohort of trustworthy, well-governed organisations¹¹⁶. Greater risks also necessitates a more rigorous approach for governing each stage of the AI lifecycle, with robust measures for consistently maintaining transparency and accountability:

The Deployment of advanced AI systems involves a four-stage approval process that begins before development has commenced. Failure to adhere to this process would mean that the system cannot be legally Distributed. This approval process is based upon a series of highly endorsed recommendations for safe AGI governance^{117, 118, 119, 120}.

- A. **Stage 1 - Pre-Development:** the AI Lab must begin the application process by submitting the following documentation to prove that they meet the preliminary standards for the safe development of Advanced AI systems. This would include:
 - a. **Project Plan** for development and deployment: including timelines, objectives, accountable individuals, and criteria for progressing between stages.
 - b. **Safety-critical architecture** summary: including predicted capabilities, reliability and explainability of the system, with evidence demonstrating how the proposed level of safety will be achieved.
 - c. **Risk mitigation strategy:** a detailed explanation of how technical and operational risks will be managed, including cyberattacks, leaks/espionage, conflicts of interest, post-deployment misuse and rogue behaviour.
- B. **Stage 2 - Post Development:** once development is approved and completed, the AI Lab must then assess the safety of the AI system, and report their findings to the Australian regulator. This safety assessment must include:
 - a. Gap analysis between predicted capabilities and safety, and the results which were actually achieved by the AI system.
 - b. Red-teaming protocols¹²¹, including a comprehensive assessment of dangerous capabilities, misalignment and potential for misuse¹²².

¹¹⁶ Hendrycks et al. (2022) Unsolved Problems in ML Safety; <https://arxiv.org/abs/2109.13916>

¹¹⁷ Anderljung & Hazell (2023) Protecting Society from AI Misuse: When are Restrictions on Capabilities Warranted? <https://arxiv.org/pdf/2303.09377.pdf>

¹¹⁸ Shevlane, T (2022) Structured access: an emerging paradigm for safe AI deployment; <https://arxiv.org/pdf/2201.05159.pdf>

¹¹⁹ Anderljung et al. (2023) Frontier AI Regulation: Managing Emerging Risks to Public Safety; <https://arxiv.org/pdf/2307.03718.pdf>

¹²⁰ Schuett et al. (2023) Towards Best Practices in AGI Safety and Governance; governance.ai

¹²¹ Ganguli et al. (2022) Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned; <https://arxiv.org/abs/2209.07858>

¹²² Shevlane et al. (2023) Model Evaluations for Extreme Risks; <https://arxiv.org/abs/2305.15324>

- c. Third-party safety audits, as per Raji et al. (2022)¹²³ and Mokander et al. (2023)¹²⁴ performed by at least two internationally-recognised organisations (equivalent to the Alignment Research Centre¹²⁵). To avoid conflicts of interest, these organisations should eventually be assigned by an international governing body¹²⁶, similar to the one recently proposed by the UN¹²⁷.
 - d. Cybersecurity penetration-testing via a third-party provider; should include attacks including data theft and service disruption¹²⁸.
- C. **Stage 3 - Pre Deployment:** once the post-development analysis has been undertaken and approved, the AI Lab may then prepare for Deployment. This includes:
- a. Providing evidence that technical solutions have been found for dangerous capabilities or other risks identified in Stage 2; this should include updates to safety-critical architecture and follow-up assessments by third-party auditors.
 - b. An explanation of technical and governance infrastructure that has been developed to monitor for post-deployment risks (e.g. misuse, rogue behaviour)¹²⁹.
 - c. An incident response strategy for responding to post-deployment risks once they are detected.
 - d. Evidence of having addressed cybersecurity vulnerabilities revealed in Stage 2, where applicable.
- D. **Stage 4 - Post deployment:** if the AI Lab proceeds through all 3 stages, they may then progress to deploying their advanced AI system for use in Australia. Post-deployment risk management leverages a specific communication arrangement between Users, AI Labs, and Regulators. This is described in the following:
- a. The Developer transfers the AI system to a high-security data centre in Australia.
 - b. Once secured, the system then can be accessed via a Government operated API¹³⁰, which records metadata about usage of the system (see Part 4).
 - c. The inputs and outputs of the model are encrypted and transferred to the AI Lab, who is responsible for detecting and reporting any suspicious activity.
 - d. Users must also monitor and record all inputs and outputs during their usage of advanced AI systems; these are used as part of an ongoing their risk-management program (see Part 4).

¹²³ Raji et al. (2022) Outsider Oversight: Designing a Third Party Audit Ecosystem for AI Governance; <https://arxiv.org/abs/2206.04737>

¹²⁴ Mokander et al. (2023) Auditing Large Language Models: A Three Layered Approach; <https://arxiv.org/pdf/2302.08500.pdf>

¹²⁵ Alignment Research Centre (2023) Update on ARC's recent eval efforts <https://evals.alignment.org/blog/2023-03-18-update-on-recent-evals/>

¹²⁶ Ho et al. (2023) International Institutions for Advanced AI; <https://arxiv.org/abs/2307.04699>

¹²⁷ Quach (2023) UN boss recommends nuclear option for AI regulation; [The Register](#)

¹²⁸ Heim & Ladish (2023) *Information security considerations for AI and the long term future*; [Heim.xyz](#)

¹²⁹ OpenAI (2022) Lessons learned on language model safety and misuse; <https://openai.com/research/language-model-safety-and-misuse>

¹³⁰ Shevlane, T (2022) Structured access: an emerging paradigm for safe AI deployment; <https://arxiv.org/pdf/2201.05159.pdf>

Note: some existing AI Labs already have the motivation and capabilities to meet the requirements for Deploying systems in accordance with these requirements. These are sometimes referred to as “AGI labs”, which have combined expertise in AI safety and advanced AI systems. Examples include Anthropic, Google DeepMind, and OpenAI.

Other organisations with explicit capabilities or intentions to develop advanced AI systems, but without well-demonstrated risk-management capabilities, include organisations such as Baidu, Meta, Microsoft and Nvidia.

- E. Regulators must coordinate with international authorities to detect and prevent unauthorised Deployment of advanced AI systems. Any attempt to do so should be considered by the Australian Government to be a criminal offence under the Criminal Code Act; i.e. an example of “creation and distribution of malicious software”¹³¹.
 - a. The enforcement of these laws will require relevant government agencies, such as the Australian Signals Directorate, AFP, DFAT and ASIO, to develop sufficient capabilities in order to detect, respond to and prevent unauthorised deployment of advanced AI systems, both domestically and internationally.
 - b. This will include activities such as monitoring usage patterns of utilities such as electricity, internet and cloud computing, distribution of high-performance computer chips, while also cross checking this data against other potential predictors such as frequency of social media activity, emails or asset purchases.
 - c. For examples of what to look for, some open-source¹³² and free¹³³ products already border on the distinction between “precursor” and “advanced” AI systems.

Part 4: Use of Advanced Systems

Once an advanced AI system is safely deployed in Australia, User access can then be provided to Australian organisations. This access is granted using a permit scheme, similar to precursor systems, although including a number of additional risk management measures.

- A. In order to obtain a permit, a User must first provide evidence and documentation in support of their request. This should include:
 - a. Detailed description of the project or program that the Advanced System will operate within. Must specify activities and tasks that are within scope for the Advanced system to carry out.
 - b. Evidence of sufficient reason to need advanced systems; must prove that narrow AI or precursor systems are thoroughly inadequate for the proposed initiative.
 - c. A comprehensive analysis of how the outputs meet the legislated criteria for what an Advanced Systems can be used for.

¹³¹ Cybercrime law; [AFP.gov.au](https://www.afp.gov.au)

¹³² Significant Gravititas (2023) Auto-GPT: An Autonomous GPT-4 Experiment; <https://perma.cc/2TT2-VQE8>.

¹³³ AgentGPT. <https://agentgpt.reworkd.ai/>

- d. Proof of adequate governance structures and technical expertise to manage the system safely.
 - e. Technical documentation outlining how the usage and outputs of the system will be monitored and securely stored in preparation for audits.
 - f. A comprehensive risk management strategy, including incident response protocols, and accountable stakeholders within the organisation.
- B. If an organisation's application is approved, they will then be subject to compliance checks in the form of random audits. These audits will involve reviewing data on usage and outputs of the systems, as well as auxiliary assessments on Governance structures and project outcomes. There are various forms of noncompliance, including:
 - a. Using an advanced system for tasks other than those which were originally approved in the application process.
 - b. A failure to monitor down-stream effects or externalities caused by the system.
 - c. Failing to recognise and act upon dangerous or suspicious behaviour.
 - d. Evidence of tampering with the data, or using the system in ways that make the effects of the outputs unclear.
 - e. Evidence that the contributions made by the advanced AI system are not serving the public benefit to the degree which was expected when the project was approved.
- C. Audits are carried out by a centralised governmental auditing body (such as an AI Commission). In instances where the system is found to be noncompliant, either or both of the organisation (user) and the provider (system developer) can be penalised. Such instances include:
 - a. Unlawful or negligent use of the system, where reasonable safeguards should have prevented the system from undertaking the activity in the first place; in this case both entities are penalised.
 - b. Noncompliance due to a fault in the system (e.g. deceptive or unsafe behaviour) occurring without being identified or addressed. In this case, both the organisation (user) and the Developer face penalties.
 - c. The system is used for purposes outside the scope of the agreed terms. In this case, only the organisation (user) is penalised.