Australian Government | AUSTRALIA'S CHIEF SCIENTIST

Rapid Response Information Report

Generative AI: Language models and multimodal foundation models

24 March 2023

This report was commissioned by Australia's National Science and Technology Council at the request of the Minister for Industry and Science, the Hon Ed Husic MP in February 2023.

The questions asked were:

- What are the opportunities and risks of applying large language models (LLMs) and multimodal foundation models (MFMs) learning technologies over the next two, five and ten years?
- What are some examples of strategies that have been put in place internationally by other advanced economies since the launch of models like ChatGPT to address the potential opportunities and impacts of artificial intelligence (AI)?

Australia's National Science and Technology Council is responsible for providing advice to the Prime Minister and other Ministers on important science and technology issues facing Australia. The Council is Chaired by the Prime Minister, with the Minister for Industry and Science as the Deputy Chair, and Australia's Chief Scientist as Executive Officer.

This report was convened by the Australian Council of Learned Academies, alongside the Australian Academy of Humanities, the Australian Academy of Technological Sciences and Engineering and the Australian Academy of Science. Lead authors were Professors Genevieve Bell AO FTSE FAHA, Jean Burgess FAHA, Julian Thomas FAHA and Shazia Sadiq FTSE.

Rapid response information reports provide a synthesis of the available scientific and technical information at the time commissioned, and expert scientific opinion, and are peer reviewed by experts in the field. Rapid response reports typically do not provide recommendations or policy advice to government.

This report represents independent science evidence to government, and does not represent the views or policies of the Australian Government or Australia's Chief Scientist.

To cite this report:

Bell, G., Burgess, J., Thomas, J., and Sadiq, S. (2023, March 24). Rapid Response Information Report: Generative AI - language models (LLMs) and multimodal foundation models (MFMs). Australian Council of Learned Academies.

Rapid Response Information Report: Generative AI - language models (LLMs) and multimodal foundation models (MFMs)

24 March 2023

This rapid research report addresses the questions:

- **What are the opportunities and risks of applying large language models (LLMs) and multimodal foundation models (MFMs) learning technologies over the next two, five and ten years?**

- **What are some examples of strategies that have been put in place internationally by other advanced economies since the launch of models like ChatGPT to address the potential opportunities and impacts of artificial intelligence (AI)?**

Summary points

- ChatGPT is an early example of the kinds of applications and services that will emerge from Generative AI built on LLMs and MFMs. It was developed by the American AI organisation OpenAI – a combined non-profit/for profit entity that received a U$10B investment from Microsoft in 2023. ChatGPT is based on a large language model (LLM) and uses considerable pre- and post-processing of data to deliver a compelling user experience.
- Given the speed of innovation, quantum of investment and lack of technical information, it is almost impossible to accurately forecast opportunities over the next decade. Known risks are clearer, but there are categories of emerging risks that are difficult to forecast. In the shorter term, generative AI, based on LLMs and MFMs, will likely impact everything from banking and finance to public services, education and creative industries.
- Generative AI will raise questions about opportunities and risks of widespread adoption; the scope and adequacy of national strategic planning and policies; the fitness of legal and regulatory approaches; and the implications of increasing geo-political competition and geo-specific regulations.
- Generative AI presents opportunities across various industries, including healthcare where LLMs and MFMs are being used to analyse medical images and consolidate patient data, and in engineering to evaluate and optimise designs.
- The current concentration of generative AI activities poses risks for Australia and raises questions about our capabilities, capacities, investments and regulatory frames. Questions include: do we have sufficient compute power, appropriately skilled practitioners, scientific expertise, workforce development strategies and policy settings that range from critical technologies, to education, ethics, governance and regulation?
- LLM and MFMs are generating a surge in interest, innovation and investment. Much of this work is happening inside commercial organisations, and is currently concentrated in a small number of organisations and countries, notably the US and China.
- While some of the architectural innovations are being shared publicly, overall there is a paucity of information about the development, deployment and commercialisation of these models and the applications and services based upon them.

The current 'ChatGPT moment' is provoking public conversation about the role AI should have in Australian society. This report has been written in the context of rapid change in the ecosystem and heightened expectations about both the risks and possibilities of both ChatGPT in particular and generative AI more broadly.

Generative AI raises questions about opportunities and risks of widespread adoption; the scope and adequacy of national strategic planning and policies; the fitness of legal and regulatory approaches; and the implications of increasing geopolitical competition and geo-specific regulation in AI-related technologies and industries.

This report explains how generative AI, based on LLMs and MFMs, currently works, given that the technologies are nascent and rapidly evolving (e.g., GPT-4 was publicly released on 14 March 2023 with some multi-modal input functionality and Baidu released Ernie Bot on 16 March with multi-modal output functionality)[1] as are the business models, applications and services that are built upon them. Against this backdrop, the report explores foreseeable risks and opportunities, based on current patterns of uptake and application.

Defining generative AI

Whereas conventional AI has been largely analytic, generative AI takes its name from its capacity to generate novel content, as varied as text, image, music and computing code, in response to a user prompt. For example, conventional AI can be used to analyse features of a legal contract, such as to identify whether the contract deals with intellectual property or privacy. By contrast, generative AI can be used to generate (i.e. draft) a new legal contract to cover those issues.

GPT-3 (Generative Pre-Trained Transformer 3), which powers the free version of ChatGPT[a], is an example of new generative AI, built on an LLM. The launch of ChatGPT (a generative AI-powered chatbot) in November 2022, by OpenAI, has prompted an extraordinary amount of activity – from adversarial exploitation and forensic testing to better understand how the system works and its governing rules, to creative exploration. ChatGPT is more consumer-friendly than prior AI systems and it has been fundamentally misunderstood, from attributing it sentience to claiming that it is thoughtfully summarising the internet. Like earlier generations of AI, generative AI relies on complex mathematical models, considerable computing power and extensive human resources to train, develop and deploy.[2]

Large language models (LLMs) and multimodal foundation models (MFMs)

First developed in the 2010s, LLMs and MFMs use sophisticated machine learning algorithms to predict an output – such as an image or word – based on an input, such as a sequence of words. What all these models do is recognise patterns in data and produce sophisticated answers based on those patterns. The models are not intelligent or able to necessarily determine fact from fiction in their inputs or training data.

LLMs specialise in generating human-like text by training on vast quantities of text[b].[3] MFMs are more complex as they use a wider range of information, including images, speech, numerical inputs and code,[4] and they are trained on the relationship between the various inputs. Like LLMs, MFMs generate output

[a] ChatGPT Plus, a paid version of the chatbot uses GPT-4. The free version uses an older version, GPT-3.5.

[b] GPT-3 is trained on about 45TB of text data from different datasets, including Wikipedia and books. GPT-3's training data is known to only contain information up to September 2021.[145] For one GPU, it would take over 300 years to train the model, and it cost 5 million dollars to train the neural network.[146]

based on learned patterns from the training input. Like LLMs, they also require tremendous computational power to train their models.[c]

The predictive text and image generation functionality of LLMs and MFMs is not new. What is novel is the scale of the data used for training and the extremely large number of parameters in the models. Recent advances in architecture and modelling have made it possible to dramatically increase the size of datasets on which LLMs and MFMs can be trained, and as a result the ability to ascertain a richer set of contextual patterns and probabilistic relationships between data. For example, GPT-4 will be trained on 100 trillion parameters whereas GPT-3 is currently trained on 175 billion. This scale is critical in allowing the models to account for the input context in a more nuanced way.[3] Because of this, LLMs and MFMs are now far more powerful than their predecessors.

That said, most people are not directly encountering LLMs or MFMs; rather, they are encountering new kinds of services, applications and businesses that use them, whether in the form of chat-bots, enhanced applications or subscription services. For example, ChatGPT provides a seamless user experience in both requesting and receiving information from a LLM (see Figure 1). Invisible to the user, ChatGPT uses pre-processing and post-processing to calibrate whether the prompt is appropriate and return the answer in a form that seemingly responds to the original request. Developers have decided what and how user requests for information should be handled, including whether or not to label the request appropriate or ethical based on OpenAI's internal guidelines. It is not yet clear what the implications of such choices might be for third parties who integrate these services and applications into their own ecosystems.

The LLM/MFM lifecycle

To date, successfully developing an LLM or MFM has required substantial monetary, computational and human resources.[5] The data required, the processing power needed, and the risks and potential consequences of 'wrong' answers or malicious uses amplify these challenges. As a result, developers are employing an evolving range of strategies to design systems in ways that may prevent social harms (e.g., inequities, misinformation), maximise user safety (e.g., protecting vulnerable communities), and/or maintain some degree of control over downstream applications (see Table 1).[6]

[c] GPT-4 appears to be an MFM that can process documents that contain both text and image data, answering detailed questions about images in plain text. Due to OpenAI's most recent policy on relaying information about their models (as seen in their preprint OpenAI 2023), we do not know the size of the model, its internal structure or the dataset that OpenAI used to create GPT-4, nor do we have an estimate about the computational power used to create this model or the cost.

Example LLM user experience

(based on ChatGPT-3)

What the user sees

Visual interface

Prompt

Response

Prompt

Response

What ChatGPT does

ChatGPT selects its responses from a pre-trained Large Language Model (LLM).

An LLM is an AI designed to understand and generate human-like language.

The current model (GPT3.5) has 175 billion parameters and three billion words.

The ChatGPT application shapes its output based on pre-determined rules and previous interactions.

Typical ChatGPT user experience

A user initiates a conversation with ChatGPT by typing a message or question into a chat interface, such as a messaging app or a chatbot on a website.

ChatGPT receives the user's message.

ChatGPT uses its natural language processing capabilities to predict the user's intent from the context of the conversation.

ChatGPT generates a response based on its training and the user's input.

The response is displayed in the chat interface, and the user can respond with follow-up questions or feedback.

The conversation between the user and ChatGPT continues in this way, refining future responses based on the information provided.

Chat GPT3.5 operating process

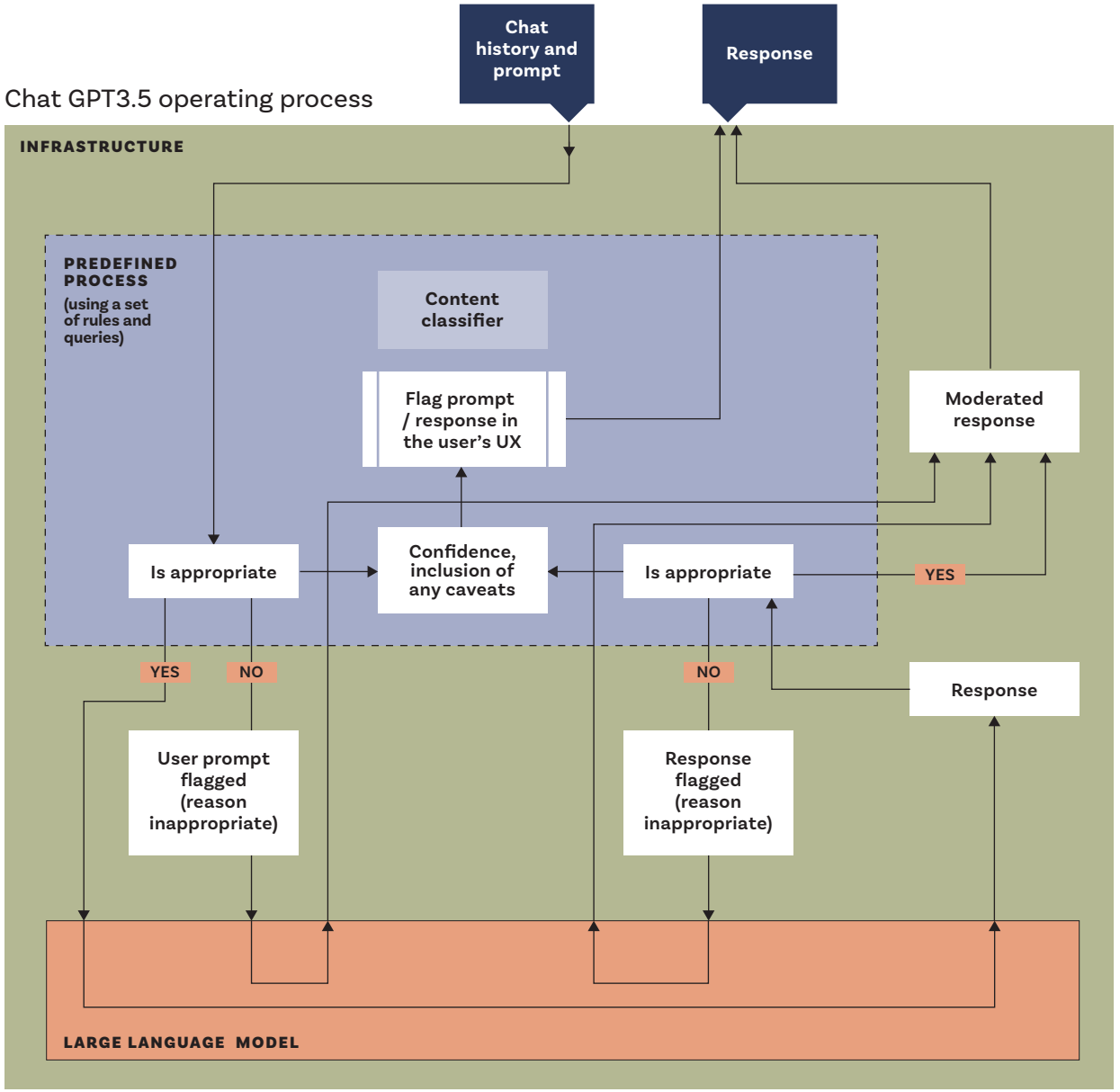


Table 1: Activities and risk management strategies throughout the generative AI model development lifecycle *(source: developed for this rapid response report, on Generative AI, to NSTC)*

Lifecycle Stage	Activity	Risk management strategies
Model pre-training	Models are trained on large, general datasets to perform general, abstract tasks like predicting subsequent words (LLMs) or associating existing images with captions (MFMs).	Data are edited, added or removed to reduce bias and improve quality, and sensitive or undesired data or attributes may be removed (e.g., faces, adult content, copyright images).[6–8]
Model fine-tuning	After pre-training on general data, models are fine-tuned for a specific application that builds on the general abilities learned in pre-training, such as helpfully and safely answering user queries (LLMs), or generating novel images from descriptions (MFMs). Developing these models require more human intervention.	Fine-tuning relies on human judgement; it involves the use of static datasets (supervised learning) or interactive feedback from humans or other automated tools (reinforcement learning).[5,7–9] Fine-tuning is critical to make LLMs or MFMs safer and more useful, e.g., ChatGPT was fine-tuned to become a dialog agent that can answer queries in a prompt-response format and follow instructions.
Implementing input and output filtering	Developers shape the user inputs to a model, and the generated outputs from a model, to mitigate risks or further improve model performance on the target application. Developers try to balance allowing users to 'steer' a model's output to specific applications, while also enforcing 'guardrails' that prevent malicious or out-of-scope uses.[5]	Inappropriate or out-of-scope user inputs (e.g., requests to generate adult content or medical advice) may be blocked or receive special treatment using, e.g., keyword matching. Model outputs may also be restricted using existing content moderation APIs or bespoke interventions (e.g., to prevent generation of adult content, or of celebrity deep fakes).[10,11] Outputs may also be watermarked to highlight that the content was AI-generated.[12]
Pre-release testing	Developers typically test their application throughout the development process, and especially before release, to identify and amend problems. Due to the higher stakes, model testing is becoming more planned, structured and intentional.	'Red-teaming', a technique used in cybersecurity, is employed – whereby a group of users try to find flaws in the system and/or suggest design improvements.[11,13,1412] 'Fuzzing' may also be used, whereby automated tools change system inputs and check for incorrect outputs.[15]
Release and distribution	The developer of an application can choose a range of strategies for making the system available. Historically, AI researchers have tended to open-source their model source code and parameters, but as LLMs and MFMs increase in cost and potential risks, release strategies are becoming more controlled.[16–18]	Limited access protocols (e.g., giving developers paid access to an API and the general public access to a web interface (which OpenAI does for ChatGPT) may be implemented, rather than publishing a system as open-source. Specialised licences for datasets and models are also increasingly employed.[19] Model documentation may also be released, and there are efforts to standardise the information that is reported (e.g., features, risk, limitations, assumptions).[20–22]
Post-release monitoring	After release, a developer will continue to monitor the way users interact with an application to make appropriate updates.	Developers have in the past offered monetary rewards or priority access to advanced models for users who report problems such as biases or bugs or who contribute to evaluating model performance.[23–25] Internal and external (e.g., academics or public interest groups) auditing may also be done to examine the design of models to find potential harmful effects.[26] In the worst case, a developer may even attempt to partially or fully retract an LLM or MFM application that is found to be deeply problematic.[27]

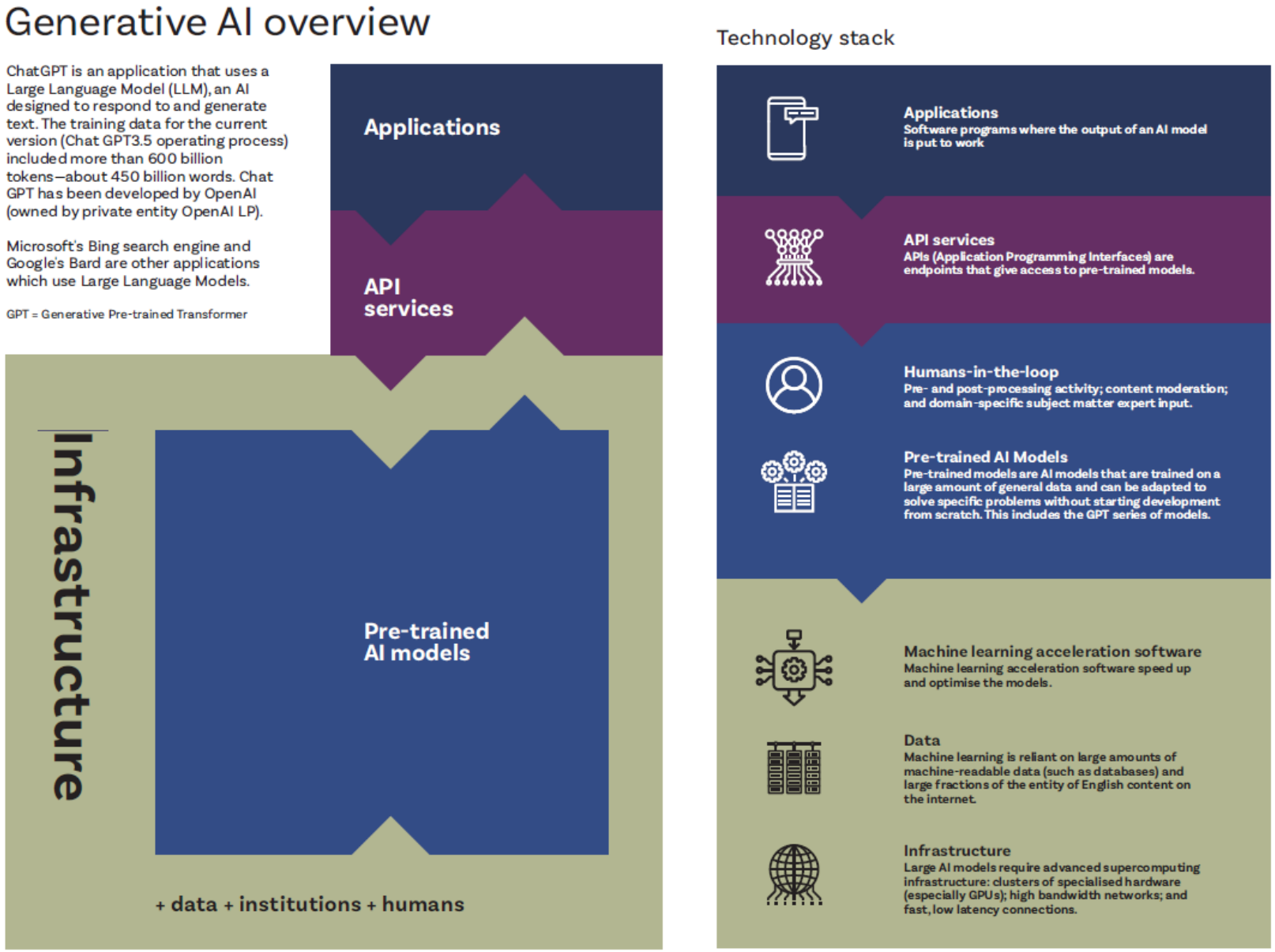
Who is developing LLMs and MFMs?

To date, despite OpenAI catalysing the most recent interest in generative AI, big companies such as Alphabet[d] and Microsoft are at the forefront of LLM development and their initial monetisation. Recently, Chinese and Indian tech companies have announced their own LLM-based chatbots, including Baidu's Ernie Bot. [28–30] Meta has also announced its own LLM (LLaMa) aimed at academics, with a more energy efficient footprint. In pre-print articles, the developers of LLaMa have shared their dataset sources for their model and their fine tuning, as well as their entire model's architecture.[31,32] Some universities, such as Stanford, are also innovating in this space, building on LLaMa, to create an open source model trained on ChatGPT, running on a laptop, for less than U$600.[33] If this work proves stable and scalable, it has significant implications for the current eco-system and its business models.

LLMs and MFMs, and the applications, services and business models that are built on them, rely on a larger technology 'stack', including application program interfaces (APIs), machine learning operation management (MLOps), machine learning (ML) acceleration software, and supercomputing and cluster-based infrastructure (see Figure 2). This means there are lots of other players involved in the current generative AI wave. As more services, applications and business are built on top of LLM/MFMs, regulation may be necessary for safe and responsible management of generative AI and there are multiple points of intervention possible within the technology stack (see Figure 2).

[d] Google's addition (Bard) that was released on 21 March 2023.

Figure 2: The generative AI technology stack *(source: developed for this rapid response report, on Generative AI, to NSTC)*



The concentration of generative AI resources within a small number of large multinational and primarily US-based technology companies poses potentials risks to Australia. Given the resource-intensive nature of LLM/MFM-based generative AI, can Australia be competitive in the production or fundamental research of these technologies? Australia has capability in AI-related areas like computer vision and robotics, and the social and governance aspects of AI, but its core fundamental capacity in LLMs and related areas is relatively weak.[34] While the Australian Government announced investments of $100 million in AI-related initiatives (e.g including a national AI centre),[35,36] creating generative AI technologies has especially high barriers to access, due to its considerable compute and data requirements. The technologies also have requirements for skilled workers, OpenAI team currently has 375 employees.[37] We do have large public datasets, such as the Bureau of Meteorology, Australian Bureau of Statistics and the National Library of Australia's Trove, that could offer opportunities for generative AI in Australia.

Opportunities for LLM and MFMs

LLMs and MFMs, and the applications, services and business models based on them, have implications for the Australian economy, now and into the future. It has long been recognised that AI-based technologies may lead to job losses where they enable machines to replace the work of humans undertaking particular tasks; a recent analysis of the exposure of U.S. workers to the potential effects of GPTs estimates that in 19% of U.S. jobs, at least 50% of tasks are exposed, meaning that these tasks correspond to the known capabilities of generative AI. The jobs concerned span all income levels.[38] In a broader study of the economic impacts of AI, there is the potential for excessive automation, where firms do not take negative impacts on workers into account.[39]

Others observe that automation often complements human labour, such as workplaces where robots are introduced often increasing total employment.[40] The fact that a machine may perform one or more relevant tasks does not mean that job replacement will necessarily occur, indeed the strongest business cases for investment in AI are likely to emphasise the creation of additional value to products or services rather than savings in labour costs.[41] In these cases, technologies such as generative AI are likely to both create new jobs and augment existing ones by enhancing human decision-making skills.[41,42]

The World Economic Forum's *2020 Future of Jobs* Report combines statistical evidence with interviews with business leaders in order to describe the future outlook for jobs and skills. The Forum predicts that currently emerging technologies, including but not limited to generative AI, will create more jobs than they destroy, but notes that while job destruction is accelerating, job creation is slowing.[43] Policy makers should be prepared for the challenge of balancing investment in generative AI with the need to manage job displacement and ensure decent human working conditions in automated workplaces.[44,45,46]

As with the introduction of any new technologies, the impacts are likely to be unevenly distributed.[47]An initial impact on workplace productivity may involve the automation of basic tasks, such as letter-writing, and supporting higher-level tasks, for example, web development. Early adopters of existing AI, including information-based industries, research and healthcare, may benefit, through, for example, textual analysis and image processing, before a wider uptake of the technology occurs across the economy.

To maximise economic benefit, businesses will need to integrate generative AI into their operations and develop new business models, products and services, in addition to using the new tools to enhance the productivity of existing processes. They will also need to contemplate new job categories and skills – for instance, data curation and prompt engineering. There could be an increase in demand for digital skills, with the entry of generative AI expected to require 161,000 AI specialisations globally by 2030.[48,49] It is predicted that there will be a rapidly increasing demand for skills in digital technologies in Australia, with skills in software orchestration/automation, AI and data analysis predicted to grow especially quickly. However, some of these new occupations, such as data scientist and data engineer, do not exist in the current Australian occupation classification system.[50]

Measures of the Australian digital economy, including the mapping and classification of the skills it will require, are not yet well developed. However, a substantial and capable workforce will be critical for Australia to be generative AI-ready. Over the next two to five years especially, competition for talent both domestically and internationally and a shortage of skilled workers represent a key risk for Australia.[e]

---

[e] A cautionary example is provided by the United States, in which the tech sector has recruited many teaching and research ML specialists from the university sector, leaving few staff left to teach and supervise PhDs. This has led to a

Future industry and public sector opportunities

The rapid expansion of ChatGPT shows the potential of generative AI technologies is difficult to predict over the next two years, let alone ten. This is especially true given the large financial investments that are currently being made, and the expectations of future profits. However, one way to model what might happen is to focus on impact spaces rather than specific opportunities.

LLMs and MFMs are already being used across a range of industry settings from health and engineering to social services and creative industries. As noted elsewhere, we are also seeing these models being integrated into existing systems, including search and productivity software such as Microsoft 365 Copilot, Google Workspace and Bing.[51–54]

LLM and MFMs are being used to analyse medical images [55–57] and consolidate patient data,[58,59] evaluate and optimise engineering designs,[3] support social service provision,[60] analyse and generate documentation in legal services, generate creative material in the visual arts, music and filmmaking, as well as journalism, advertising and marketing.[61,62] Educators are experimenting with the integration of LLMs and MFMs into classroom activities, curriculum and assessment, as well as developing services to detect their inappropriate use.[63]

These current applications are an indication of the very broad spectrum and potentially rapid uptake of future integrations into industries and the public sector. As these applications develop further it will be important to understand how they interact across sectors; how applications based on LLMs and MFMs integrate and connect with each other, and how they interact with existing general-purpose technologies, such as search engines.

Risks of LLMs and MFMs

LLMs and MFMs, and the services, applications and businesses built with them, have already amplified longstanding public and expert concerns about the higher-scale risks of AI, including existential risks.[64] For instance, conversations about ChatGPT, in daily life and in the press, routinely evoke questions about what it means to be human, the role of computing in daily life, the perils of next-stage automation and fears about runaway, uncontrollable technology.[65]

Heightened concerns could create a polarised and unproductive public debate, which may then dominate our responses to future uses of these applications. There is unlikely to be a consensus on these issues. It will be important to maintain an active and informed conversation on the uses and applications of these emerging technologies.

While LLM and MFM-based generative AI is relatively new, and the services, applications and businesses utilising them are nascent, we can build on what we already know to make sense of the risks posed on a spectrum from narrow to broad categories. There are three important categories of risk:

- **Technical system risks**, both for the model itself and its data. These include validity and reliability; trust in and accuracy of answers; safety; security and resilience; system accountability and transparency; explainability and interpretability; privacy; management of biases and other assorted quality assurance considerations.

---

critical undersupply of ML-qualified graduates.[147,148] Joint positions between industry and academia are being explored as a potential solution to this problem.[79]

- **Contextual and social risks**, including risks to human rights and values arising from AI use in high-stakes contexts (such as law enforcement, health and social services), and risks posed by the more 'routine' deployment of AI that reproduces and accelerates existing social inequalities.
- **Systemic social and economic risks**, including impacts on democratic systems; social discourse and dialogue; environmental impacts; transformation of work; mistrust in private and public sector organisations and market dominance by a small number of transnational corporations providing generative AI as a platform or service (including the issues of maintenance, and decommissioning of legacy systems).

The extent that these risks are realised or mitigated will depend on the actions of governments, industries, developers and consumers. Trust will be critical for the adoption of LLMs and MFMs and the applications and services built on them. Trustworthiness will be built through appropriate levels of reliability, transparency, accountability and legal, policy and other safeguards.

Generative AI will also raise regulation and deployment considerations to ensure existing and new inequalities are not exacerbated or initiated. AI tools require considerable internet bandwidth, power and suitable devices, which are not available or affordable to everyone. Regional Australians and older Australians particularly experience poorer digital inclusion.[66]

Accuracy and bias

A major limitation of LLMs, and the applications built on them, is the accuracy and quality of the answers generated. They are only as good as the data they are trained on; the models use statistical analysis to determine the 'correct' next word, not an understanding of the content, and the user interfaces can shape the way users perceive the validity of the answers.

Inaccuracies and bias

- In some cases, outputs can be entirely erroneous, or simply misleading – known as 'hallucination'.[67] Furthermore, the way a question is asked (to ChatGPT, for example) changes the perceived tone of confidence in the response. For example, if the user includes words such as 'expert', 'technical' or 'consultant' in their question, the system may respond with 'experts say' prefacing incorrect information.[68] Future generations of LLMs may need to cite genuine sources to provide sufficient reasoning for their results – noting that currently they sometimes invent references when asked, with potential problematic impacts.[69]
- Representational bias where, for example, a model is only trained on Western literature or male 'voices', can exacerbate existing social inequalities.[70,71] The consequences could be severe if applied to sectors such as law enforcement, recruitment (e.g., translation services) and social services. [72]
- While healthcare is a key opportunity, it provides a case study for how a sector can replicate existing biases. The exclusion of women, and other miniority groups, from medical research is well documented, leading to poorer health outcomes.[73] Conversely, well-designed LLM tools could assist in countering medical biases with a more reliable assessment of reported symptoms. Finally, models trained on overseas datasets may fail to capture place-based factors, such as diagnostics for bushfire-related respiratory issues. Deliberate training and review can address these limitations.[74]
- In predictive policing, one example of 'dark forecasting', AI can perpetuate existing inequalities in over-policed populations. Some previous AI-predictive policing tools have been discontinued due to inbuilt bias.[75,76] Racial profiling has been the subject of scrutiny as this can be both countered and inflated by algorithms.[77]

Misinformation

- LLMs and MFMs have the potential for misuse by generating high-quality, cheap and personalised content, including for harmful purposes. Tools built on these models are already in use to generate deep fakes (high-quality artificial images, video and speech for disinformation, including by state actors) indistinguishable, at least without special training or access to technical tools, from human-generated content. [78] Existing challenges related to the spread of misinformation may be amplified as AI-generated content circulates alongside other information.
- LLM-generated content could also be misused in democratic processes such as parliamentary consultations by creating a flood of submissions to mislead public opinion. While they provide ample opportunities for misuse, the capability of generative AI can also be used to detect harmful content, as well as the inappropriate use of generative AI in other contexts, such as education settings.[79]

Pre- and post-Processing

- Applications and services based on LLMs and MFMs use a range of pre- and post-processing activities, such as restating the query in a way that the LLM or MFM can best address, and declining 'inappropriate' requests based on pre-defined rules. However, this processing encodes values that are not always transparent to the user or potential regulators, and can be subverted through adversarial practices such as 'jailbreaking'[80] - the process of exploiting a system's features to remove provider-imposed restrictions on its use.

Human rights

Where an AI-enabled system has no clear human decision-makers, it is challenging – but essential – to establish responsibility for adverse impacts.[47] Most LLMs and MFMs are 'black box technologies' where the public cannot understand how the model arrives at its outputs, making it difficult, or potentially impossible, for a human to assess the reliability of the results or seek redress.[58]

Institutional protections that apply a 'human-in-the-loop' approach to ensure accountability and fairness may assist, alongside other design considerations, in addressing these issues for future digital services.'Human-in-the-loop' requirements may not be appropriate where the benefits of an application are dependent on efficiency at scale; some risks will be better addressed by other approaches to monitor risks, identify errors, and provide access to remedies. Comprehensive and ongoing risk assessments and human rights due-diligence may help identify risks and mitigation strategies that are context-sensitive and appropriately tailored.[81]

Data privacy, security and sovereignty

To date, commercial organisations building LLMs and MFMs have not shared a great deal of specific details regarding the training datasets they are using, and their providence, which could include the purchase of third-party datasets and data scraping. It seems likely that permissions have not always been provided for use of large datasets drawn from the internet. Under existing Australian privacy law regarding personal data scraping, the lawfulness of some of these training sets could be questioned. For example, the Office of the Australian Information Commissioner, via the Australian Information and Privacy Commissioner, recently made a determination that was critical of the use of data scraping by Clearview AI to build its facial recognition service. [82] Attribution and reference to licences of existing copyright material remain issues to be appropriately addressed, as is the copyright of material generated by the language models.[83] This can be seen in applications as diverse as coding and in art (both for artists whose work is used in

training data and for artists using generative AI in creative works). Data sovereignty is particularly a consideration for First Nations data.[84,85]

As systems become integrated, the management of privacy and consent when collecting, sharing and using datasets will need further attention. As generative AI is integrated into Australian systems, there will be questions regarding sovereign ownership of LLMs and MFMs, and the data they are trained on, particularly if integrated into public systems such as healthcare and education. New methods for providing and handling consent, frameworks for sharing and using data, and considerations for security in highly complex networks and with shared public–private ownership will be required.

AI presents new opportunities for data privacy breaches, for example, in the reidentification of anonymised data used for LLMs.[86] Data security is a key risk, particularly with cyberattack methods to extract training data.[87,88] In healthcare for example, there are risks for both patients who have LLMs as part of their care and patients whose data is used in a model's training dataset. This also presents an ethical challenge as patients may not have consented to this use of their data. Some say LLMs should be only trained on public data to avoid this issue; however, this results in less powerful LLMs with a higher risk of bias, as very few medical datasets are publicly available.[89]

Computing power: environmental impact and capacity

As seen in Figure 2, access to computing infrastructure is a critical enabler and challenge for some countries for generative AI. LLMs and MFMs require supercomputing-like capability, most of which is found in the US, China and Europe.[90,91]

To date, the creation of LLMs and MFMs have required large datasets stored in large data centres, and their use incurs further high costs in compute and data processing power.[92] Managing the energy and water consumption of training and retraining (including data collection and cleaning) and operating LLMs and MFMs is a challenge.[93,94] While techniques have improved the energy efficiency of algorithms, hardware upgrades and increasing levels of e-waste from computer components will heighten demand for critical minerals with resultant environmental and human rights impacts.[94,95]

International strategies to address the opportunities and risks posed by LLMs and MFMs

Internationally, general AI strategic investment has begun to focus on generative AI. Examples include the UK's AI Strategy and investment of £900 million in an AI supercomputer (to help build LLMs and MFMs), Germany's €3 billion investment by 2025, the US Government funding analysis in AI, China's plan to be a global leader in AI by 2030, and investment globally in AI start-ups.[96–98]

As the tech sector develops LLMs and MFMs, other sectors, such as finance and banking, will favour business models that leverage them. For example, Microsoft has integrated OpenAI's GPT-4 into the Bing search engine. There are indications of emerging competitive responses in this area, including the appearance of open-source development environments and an ethos of open-source AI innovation. Amazon Web Services' strategic partnership with Hugging Face, which promotes open source contributions, may signal the growing significance of alternative models for intellectual property and commercialisation in AI.[99]

Even so, the intensive infrastructural and computational resources required for the development of generative AI, and ongoing research and innovation, are concentrated in a small number of firms and countries. The trend is towards more concentration and increased geopolitical competition. For example:

- the US's CHIPS Act and parallel EU measures aim to ensure ongoing onshore computational capabilities for future AI-driven industries, with a focus on infrastructure and semiconductor design and fabrication. Initiatives such as the proposed US National Artificial Intelligence Research Resource aim to shape markets and direct innovation and competition policies towards a domestic AI innovation system more closely aligned to national interests.[100]
- China has provided policy and financial support to develop the AI industry as a national priority, with the ambition to lead in both research and application by 2030.[101–103] It has been suggested that the success of ChatGPT may be a 'Sputnik moment' for China.[104] It has coincided with a renewed focus on chips and AI research, training and recruitment, and development of Chinese language models, such as MOSS, an AI chatbot and rival to ChatGPT launched in February 2023 for public testing, and Baidu's Ernie Bot launched in March, while China has censored ChatGPT.[105,106] The result is parallel AI developments with local versions of technology developed as external versions are fenced out.

For smaller countries and markets like Australia, this competition could present challenges for access and capability, as well as the suitability of models for our context and needs. Equally, it could present opportunities for local firms, government entities and publicly funded research organisations to adapt and fine-tune small and large models for Australia-specific industries and research across all sectors.

Legal and regulatory responses

To date, global approaches to AI governance fall into two broad categories: government regulatory actions through legislation or regulator guidance, and self-regulation and voluntary standards. Whether mandatory or voluntary, these policies and regulations seek to require anyone developing and deploying AI to identify, mitigate, monitor and address risks of harm or misuse. AI specific policies and regulations operate in addition to other laws that can impact the use or misuse of AI and its applications, including privacy, tort, anti-discrimination, competition and consumer law.

Governments have been heavily involved in encouraging and supporting the development of 'soft law'[f] to allow safe but flexible innovation of AI in general.[107,108] Globally, more than 630 'soft law' AI governance programs have been identified as being developed and published between 2016 and 2019, with the number increasing substantially over time.[108] But their effectiveness is debatable. Technical standards are being developed by international standards organisations to assist in this process.[109]

Specific regulatory frameworks to address generative AI, including LLMs and MFMs, are currently being developed but have not yet been deployed in Australia or overseas. There is a growing recognition that a range of institutional measures and policies are likely to be required to mitigate public risks. However, risk management approaches are most effective within a specific context of use, and against technical rather than social or systemic risks, such as use by social media platforms. There are also challenges in regulated industries regarding auditability and what constitutes an audit trail when a generative AI is changing and adapting. Regulatory frameworks are in development for managing risks associated with AI more generally and may also be expanded to cover LLMs and MFMs.

The European Union (EU) has proposed the EU AI Act. The EU model is notable for differentiating between AI use cases: banning unacceptable uses and identifying others as 'high risk', where active ex ante compliance and ongoing monitoring is required. The model applies differentiated obligations on actors within the AI supply chain: providers, suppliers, importers and users. To date, only an EU Directive, different from the EU AI Act, tangentially deals with the question of responsibility in the case of general-purpose models.[110,111] The EU's proposed framework for governing AI may not encompass the dynamic range of contexts in which they can be used and there is debate over whether general purpose models or applications could be categorised as 'high risk'. It has been suggested that the EU AI Act, may, in the end, exclude LLMs and MFMS from its scope, until further consultation can occur.

Canada is moving in a similar direction.[112] Canada already has in place law requiring future impact assessments for the use of automated systems in the public sector. As drafted, it will apply to the deployment of systems based on LLMs/MFMs. In March 2023, Canada published plans to extend risk-based regulation and 'interoperate' with the EU Act. The initial focus of Canada's proposed new regulator, the AI and Data Commissioner, will be on education and upskilling, but the proposed legislation (the Artificial Intelligence and Data Act) is structured like the EU Act, and grants the regulator powers to require audits, and even order suspension of an AI system's use.

At the other end of the regulatory spectrum, the US relies on self-regulation, which includes public-sector driven, but voluntary multi-stakeholder processes to develop risk management and technical standards, similar processes in specific domains (such as medical devices), and contributing to international standards bodies. The US and Singapore have also developed specific tools to support AI developers and users to identify and mitigate risks.

China has its own approach involving government-led public–private sector partnerships on AI regulatory guidelines, coupled with strong government support for development of local technology and companies. A UK White Paper on regulating AI is expected shortly. Appendix 3 provides further details of approaches in different jurisdictions.

[f] "Soft law" is a term usually used to refer to statements, declarations, or sets of principles that do not have the force of law, but are still intended to influence behaviour of firms and people. In this space, it refers to documents such as the Government's AI Ethical Principles and statements and AI Ethics Principles. They have no legal force or effect (they are not mandatory; no sanction arises if they are breached) but clearly intended to influence the use and development of AI. Codes of Practice (that are not adopted into law or otherwise given legal force) can also be soft law instruments assuming they are not made binding in some way, eg via a contract.

Australia's current approach to technologies is largely through self-regulation and voluntary standards approach, but these have historically been technologically neutral. There are some laws that may impact the way AI systems are designed or the context they operate, such as the Copyright Acts 1968, Privacy Act 1988, Consumer Act 2010, Fair Work Act 2009 and laws related to anti-discrimination[g]. There is currently no legal obligation for developers to undertake a risk assessment, except firms who supply to the NSW Government (under contractual obligation rather than legislation).

Multi-stakeholder and sector-specific development

LLMs and MFMs will challenge risk-based approaches as they change the nature, predictability and scale of potential risks, and make it harder for any one entity to identify, assess or mitigate those risks.[113]

Meanwhile, governments and public and private sector organisations are responding to the specific risks ChatGPT is thought to pose. In education, for example, some Australian states, universities and schools have banned ChatGPT on the basis that it could aid academic dishonesty, while others have adopted a less restrictive approach, encouraging educators and learners to experiment, and acknowledge the service it can provide students and teaching staff.[114–117] There is public discussion too of the legal and professional risks that could arise, for example, from lawyers, ulitising ChatGPT to generate legal advice or inputting confidential information as part of user prompts.

There is a proliferation of voluntary principles, guidelines and standards for trustworthy AI and several multi-stakeholder coalitions (bringing together industry, government representatives and civil society organisations such as the OECD Global Partnership on AI and the industry-led Partnership on AI.[118] These seek to develop industry consensus around emerging best practices, including documentation such as cards for model reporting and datasheets for datasets.[119,120] Individual firms have published their own risk-based approaches: OpenAI has published 'Best Practice for Deploying Language Models';[121] Microsoft has published a range of documentation on risks,[122] threat modelling[123] and responsible use.[124] However, in 2023, we have seen the same organisations backing away from their prior guidelines and reducing their internal teams working on ethics and related issues.[125]

International standards bodies are also active in developing standards for AI risk assessment[126–129], with Australia an active participant in ISO processes. There are also some attempts at rating systems and standards being promoted by civil society and public oversight groups, such as: Ranking Digital Rights[130,131] which rates digital platform companies on human rights; and AlgorithmWatch's SustAIn, which seeks standards for social and environmental sustainability on all aspects of AI development and deployments.[95]

Areas for ongoing attention

Generative AI is transformational and is already beginning to change how we live and work. Decision-makers and the broader Australian community need a stronger understanding of its risks and opportunities if we are to successfully manage its rapid development, use and uptake over the next five years.

LLMs and MFMs are evolving very rapidly and are likely to continue to do so. At the same time, additional information and analysis about the major current models continue to appear in the public domain. A clear

[g] For example, copyright may affect the use of text and images as training data (at least if it occurred in Australia, could well infringe copyright). If personal information is involved, either as training data or inputted into generative AI systems, that would involve use of personal data subject to privacy legislation (at least if done by non-excluded entities ie large businesses and public sector entities).

understanding of the most important challenges, risks and opportunities of these models would benefit from ongoing attention. Critical areas are likely to include:

- the appearance of new LLMs and MFMs and the business models, applications and services built on them;
- critical evaluations and risk assessments of the LLMs and MFMs that help explicate the nature of training datasets, energy use, compute budgets and pre/post processing activities;
- the scale and nature of harmful social outcomes of LLM and MFM-based applications and the reasons for these outcomes;
- early examples of successful and sustained integration of LLMs and MFMs into workforce and enterprise organisations; and
- regulatory developments and responses in other jurisdictions, together with multilateral and non-government governance initiatives.

References

1. Yang, Z. China tech giant Baidu releases its answer to ChatGPT. *MIT Technology Review* (2023).
2. Lai, C., Ahmad, S., Dubinsky, D. & Maver, C. AI is harming our planet: addressing AI's staggering energy cost. *Numenta* https://www.numenta.com/blog/2022/05/24/ai-is-harming-our-planet/ (2022) [Accessed 21 March 2023].
3. Narayanan, D. *et al.* Efficient Large-Scale Language Model Training on GPU Clusters Using Megatron-LM; Efficient Large-Scale Language Model Training on GPU Clusters Using Megatron-LM. (2021) doi:10.1145/3458817.3476209.
4. Rick Merritt. What Is a Transformer Model? *NVIDIA* https://blogs.nvidia.com/blog/2022/03/25/what-is-a-transformer-model/ (2022) [Accessed 1 March 2023].
5. OpenAI. *GPT-4 Technical Report*. https://cdn.openai.com/papers/gpt-4.pdf (2023).
6. Kaminski, M. E. Regulating the Risks of AI. *Forthcoming, Boston University Law Review, Vol. 103, 2023, U of Colorado Law Legal Studies Research Paper No. 22-21* (2022) doi:10.2139/SSRN.4195066.
7. Christiano, P. F. *et al.* Deep reinforcement learning from human preferences. *Adv Neural Inf Process Syst* **2017-December**, 4300–4308 (2017).
8. § Wei, J. *et al.* Finetuned Language Models Are Zero-Shot Learners. (2021) doi:10.48550/arxiv.2109.01652.
9. § Bai, Y. *et al.* Constitutional AI: Harmlessness from AI Feedback. (2022) doi:10.48550/arxiv.2212.08073.
10. OpenAI. DALL·E Content policy. https://labs.openai.com/policies/content-policy (2022) [Accessed 2 March 2023].
11. § Rando, J. *et al.* Red-Teaming the Stable Diffusion Safety Filter. (2022) doi:10.48550/arxiv.2210.04610.
12. Aaronson, S. My AI Safety Lecture for UT Effective Altruism. *Shtetl-Optimized* https://scottaaronson.blog/?p=6823 (2023) [Accessed 2 March 2023].
13. Hugging Face. Clean Diffusion 2.0 PoC Model Card. https://huggingface.co/alfredplpl/clean-diffusion-2-0-poc [Accessed 2 March 2023].
14. OpenAI. DALL·E 2 Preview - Risks and Limitations. https://github.com/openai/dalle-2-preview/blob/main/system-card.md#external-red-teaming (2022) [Accessed 2 March 2023].
15. Gao, X., Saha, R. K., Prasad, M. R. & Roychoudhury, A. Fuzz testing based data augmentation to improve robustness of deep neural networks. *Proceedings - International Conference on Software Engineering* 1147–1158 (2020) doi:10.1145/3377811.3380415.
16. Solaiman, I. *et al.* OpenAI Report Release Strategies and the Social Impacts of Language Models. (2019).
17. Shevlane, T. Structured Access: An Emerging Paradigm for Safe AI Deployment. *The Oxford Handbook of AI Governance* (2022) doi:10.1093/OXFORDHB/9780197579329.013.39.
18. Shevlane, T. & Dafoe, A. The offense-defense balance of scientific knowledge: Does publishing AI research reduce misuse? *AIES 2020 - Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 173–179 (2020) doi:10.1145/3375627.3375815.
19. Contractor, D. *et al.* Behavioral Use Licensing for Responsible AI. *ACM International Conference Proceeding Series* 778–788 (2022) doi:10.1145/3531146.3533143.
20. Mitchell, M. *et al.* Model Cards for Model Reporting. *FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency* 220–229 (2018) doi:10.1145/3287560.3287596.
21. § Gilbert, T. K., Dean, S., Lambert, N., Zick, T. & Snoswell, A. Reward Reports for Reinforcement Learning. (2022) doi:10.48550/arxiv.2204.10817.
22. Gebru, T. *et al.* Datasheets for datasets. *Commun ACM* **64**, 86–92 (2021).
23. Garg, V. Mitigating Prompt Injection Attacks on an LLM based Customer support App. https://vaibhavgarg1982.medium.com/mitigating-prompt-injection-attacks-on-an-llm-based-customer-support-app-b34298b2bc7a (2023) [Accessed 2 March 2023].
24. OpenAI. ChatGPT Feedback Contest: Official Rules. (2022).
25. Chowdhury, R. & Williams, J. Introducing Twitter's first algorithmic bias bounty challenge. *Twitter Engineering*

https://blog.twitter.com/engineering/en_us/topics/insights/2021/algorithmic-bias-bounty-challenge (2021) [Accessed 2 March 2023].
26. Raji, I. D. *et al.* Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* 33–44 (2020) doi:10.1145/3351095.3372873.
27. Snoswell, A. J. & Burgess, J. The Galactica AI model was trained on scientific knowledge – but it spat out alarmingly plausible nonsense. *The Conversation* (2022).
28. Ahmed, N., Wahed, M. & Thompson, N. C. The growing influence of industry in AI research. *Science (1979)* **379**, 884–886 (2023).
29. Tobin, M. & Li, L. Ernie, what is censorship? China's chatbots face additional challenges. *The Washington Post* (2023).
30. Sharma, D. India gets its first ChatGPT-powered AI chatbot Lexi, here are the details. *India Today* (2023).
31. § Touvron, H. *et al.* LLaMA: Open and Efficient Foundation Language Models. (2023).
32. § Wang, Y. *et al.* Self-Instruct: Aligning Language Model with Self Generated Instructions. (2022).
33. Taori, R. *et al.* Alpaca: A Strong, Replicable Instruction-Following Model. *Stanford University* https://crfm.stanford.edu/2023/03/13/alpaca.html (2023) [Accessed 21 March 2023].
34. Desmond, M., Duesterwald, E., Isahagian, V. & Muthusamy, V. A No-Code Low-Code Paradigm for Authoring Business Automations Using Natural Language. (2022) doi:10.48550/arxiv.2207.10648.
35. Department of Industry, S. and R. Funding available for AI and Digital Capability Centres. https://www.industry.gov.au/news/funding-available-ai-and-digital-capability-centres (2022) [Accessed 21 March 2023].
36. The Hon Melissa Price MP Media Releases. $44 million to build AI and digital capability centres. *Ministers for the Department of Industry, Science and Resources* https://www.minister.industry.gov.au/ministers/price/media-releases/44-million-build-ai-and-digital-capability-centres (2022) [Accessed 21 March 2023].
37. Roose, K. How ChatGPT Kicked Off an A.I. Arms Race . *The New York Times* https://www.nytimes.com/2023/02/03/technology/chatgpt-openai-artificial-intelligence.html (2023) [Accessed 24 March 2023].
38. § Eloundou, T., Manning, S., Mishkin, P. & Rock, D. GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models. (2023).
39. Acemoglu, D. *Harms of AI*. (2021).
40. Coelli, M. B. & Borland, J. The Australian labour market and IT-enabled technological change. *SSRN Electronic Journal* (2023) doi:10.2139/SSRN.4326257.
41. Agrawal, A., Gans, J. & Goldfarb, A. *Power and Prediction: The Disruptive Economics of Artificial Intelligence*. (Harvard Business Review Press, 2022).
42. Aghion, P., Antonin, C., Bunel, S. & Jaravel, X. The Effects of Automation on Labor Demand: A Survey of the Recent Literature. *CEPR Discussion Paper No. DP16868* (2022).
43. World Economic Forum. *The Future of Jobs Report 2020*. https://www.weforum.org/reports/the-future-of-jobs-report-2020 (2020).
44. O'Neill, C., Goldenfein, J., Sadowski, J., Kelly, L. K. & Phan, T. Burnout by design? Warehouse and shipping workers pay the hidden cost of the holiday season. *The Conversation* (2021).
45. Acemoglu, D. *Harms of AI*. https://www.nber.org/papers/w29247 (2021) doi:10.3386/w29247.
46. Agrawal, A., Gans, J. & Goldfarb, A. Economic policy for artificial intelligence. *Innovation Policy and the Economy* **19**, 139–159 (2019).
47. Australian Human Rights Commission. *Human Rights and Technology*. https://tech.humanrights.gov.au/downloads?_ga=2.8631797.1991244140.1678071894-1349076290.1678071894 (2021).
48. Roos, G. & Shroff, Z. What will happen to the jobs? Technology-enabled productivity improvement – good for some, bad for others. *Labour & Industry: a journal of the social and economic relations of work* **27**, 165–192 (2017).
49. Autor, D. H. Why Are There Still So Many Jobs? The History and Future of Workplace Automation. *Journal of Economic Perspectives* **29**, 3–30 (2015).

50. Hope, A. *et al.* Digital skills in the Australian and International economies. *National Skills Commission Annual Report 2020-2021*.
51. Ahmed, M. & Haskell-Dowland, P. Google and Microsoft are bringing AI to Word, Excel, Gmail and more. It could boost productivity for us – and cybercriminals. *The Conversation* (2023).
52. Spataro, J. Introducing Microsoft 365 Copilot – your copilot for work. *Official Microsoft Blog* https://blogs.microsoft.com/blog/2023/03/16/introducing-microsoft-365-copilot-your-copilot-for-work/ (2023) [Accessed 21 March 2023].
53. Kurain, T. New AI features and tools for Google Workspace, Cloud and developers. *The Keyword* https://blog.google/technology/ai/ai-developers-google-cloud-workspace/ (2023) [Accessed 21 March 2023].
54. Mehdi, Y. Reinventing search with a new AI-powered Microsoft Bing and Edge, your copilot for the web. *Official Microsoft Blog* https://blogs.microsoft.com/blog/2023/02/07/reinventing-search-with-a-new-ai-powered-microsoft-bing-and-edge-your-copilot-for-the-web/ (2023) [Accessed 21 March 2023].
55. Kung, T. H. *et al.* Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digital Health* **2**, e0000198 (2023).
56. § Song, B., Zhou, R. & Ahmed, F. Multi-modal Machine Learning in Engineering Design: A Review and Future Directions. (2023) doi:10.48550/arxiv.2302.10909.
57. Zhang, Y. *et al.* Applying Artificial Intelligence Methods for the Estimation of Disease Incidence: The Utility of Language Models. *Front Digit Health* **2**, (2020).
58. Zhang, Y. *et al.* Applying Artificial Intelligence Methods for the Estimation of Disease Incidence: The Utility of Language Models. *Front Digit Health* **2**, 31 (2020).
59. Shen, Y. *et al.* ChatGPT and Other Large Language Models Are Double-edged Swords. *Radiology* (2023) doi:10.1148/radiol.230163.
60. Coco, B. A., Henman, P. & Sleep, L. Mapping ADM in Australian social services. (2022) doi:10.25916/XXJ4-N968.
61. Yerushalmy, J. German publisher Axel Springer says journalists could be replaced by AI. *The Guardian* (2023).
62. Lee, H.-K. Rethinking creativity: creative industries, AI and everyday creativity. *Culture & Society* **44**, 601–612 (2022).
63. Loble, L. The rise of ChatGPT shows why we need a clearer approach to technology in schools. *The Conversation* (2023).
64. Elliott, A. *The Culture of AI: Everyday Life and the Digital Revolution - 1st Edit*. (Routledge, 2019).
65. Samuel, S. The case for slowing down AI. *Vox* (2023).
66. Thomas, J. *et al. Australian Digital Inclusion Index: 2021*. https://www.digitalinclusionindex.org.au/download-reports/ (2021) doi:10.25916/phgw-b725.
67. Roller, S. *et al.* Recipes for building an open-domain chatbot. *EACL 2021 - 16th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference* 300–325 (2020) doi:10.48550/arxiv.2004.13637.
68. Oviedo-Trespalacios, O. *et al.* The Risks of Using ChatGPT to Obtain Common Safety-Related Information and Advice. *SSRN Electronic Journal* (2023) doi:10.2139/SSRN.4346827.
69. Hanff, A. ChatGPT should be considered a malevolent AI and destroyed. *The Register* (2023).
70. DeCamp, M. & Lindvall, C. Latent bias and the implementation of artificial intelligence in medicine. *Journal of the American Medical Informatics Association* **27**, 2020–2023 (2020).
71. Nadeem, M., Bethke, A. & Reddy, S. StereoSet: Measuring stereotypical bias in pretrained language models. *ACL-IJCNLP 2021 - 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Proceedings of the Conference* 5356–5371 (2021) doi:10.18653/V1/2021.ACL-LONG.416.
72. Koenecke, A. *et al.* Racial disparities in automated speech recognition. *Proc Natl Acad Sci U S A* **117**, 7684–7689 (2020).
73. Merone, L., Tsey, K., Russell, D. & Nagle, C. Sex Inequalities in Medical Research: A Systematic Scoping Review of the Literature. *Women's Health Reports* **3**, 49–59 (2022).

74. Norori, N., Hu, Q., Aellen, F. M., Faraci, F. D. & Tzovara, A. Addressing bias in big data and AI for health care: A call for open science. *Patterns* **2**, 100347 (2021).
75. Pitfalls of Predictive Policing: An Ethical Analysis - Viterbi Conversations in Ethics. https://vce.usc.edu/volume-5-issue-3/pitfalls-of-predictive-policing-an-ethical-analysis/ [Accessed 28 February 2023].
76. Spaniol, M. J. & Rowland, N. J. AI-assisted scenario generation for strategic planning. *Futures & Foresight Science* e148 (2023) doi:10.1002/FFO2.148.
77. Berk, R., Heidari, H., Jabbari, S., Kearns, M. & Roth, A. Fairness in Criminal Justice Risk Assessments: The State of the Art. *Sociol Methods Res* **50**, 3–44 (2021).
78. Kreps, S., McCain, R. M. & Brundage, M. All the News That's Fit to Fabricate: AI-Generated Text as a Tool of Media Misinformation. *Journal of Experimental Political Science* **9**, 104–117 (2022).
79. Bommasani, R. *et al.* On the Opportunities and Risks of Foundation Models. (2021) doi:10.48550/arxiv.2108.07258.
80. Rainey, C. ChatGPT jailbreak DAN makes AI break its own rules. *Fast Company* https://www.fastcompany.com/90845689/chatgpt-dan-jailbreak-violence-reddit-rules (2023) [Accessed 21 March 2023].
81. Raso, F., Hilligoss, H., Krishnamurthy, V., Bavitz, C. & Kim, L. *Artificial Intelligence & Human Rights*. https://cyber.harvard.edu/publication/2018/artificial-intelligence-human-rights (2018).
82. Office of the Australian Information Commissioner. *Commissioner initiated investigation into Clearview AI, Inc. (Privacy) [2021] AIcmr 54 (14 October 2021)*. (2021).
83. Franceschelli, G. & Musolesi, M. Copyright in generative deep learning. *Data Policy* **4**, e17 (2022).
84. Vaughan-Nichols, S. J. GitHub's Copilot faces first open source copyright lawsuit. *The Register* https://www.theregister.com/2022/11/11/githubs_copilot_opinion/ (2022) [Accessed 1 March 2023].
85. Vincent, J. The lawsuit against Microsoft, GitHub and OpenAI that could change the rules of AI copyright. *The Verge* https://www.theverge.com/2022/11/8/23446821/microsoft-openai-github-copilot-class-action-lawsuit-ai-copyright-violation-training-data (2022) [Accessed 6 March 2023].
86. Murdoch, B. Privacy and artificial intelligence: challenges for protecting health information in a new era. *BMC Med Ethics* **22**, 1–5 (2021).
87. Brown, H., Lee, K., Mireshghallah, F., Shokri, R. & Tramèr, F. What Does it Mean for a Language Model to Preserve Privacy? in *2022 ACM Conference on Fairness, Accountability, and Transparency* 2280–2292 (ACM, 2022). doi:10.1145/3531146.3534642.
88. Pan, X., Zhang, M., Ji, S. & Yang, M. Privacy Risks of General-Purpose Language Models. in *IEEE Symposium on Security and Privacy* 1314–1331 (2020).
89. Brown, H., Lee, K., Mireshghallah, F., Shokri, R. & Tramèr, F. What Does it Mean for a Language Model to Preserve Privacy? in *2022 ACM Conference on Fairness, Accountability, and Transparency* 2280–2292 (ACM, 2022). doi:10.1145/3531146.3534642.
90. Kingston AI Group. Statement by the Kingston AI Group. https://kingstonaigroup.org.au/news-and-publications/f/statement-by-the-kingston-ai-group (2023) [Accessed 8 March 2023].
91. Ghahramani, Z. *Independent Review of The Future of Compute: Final report and recommendations*. https://www.gov.uk/government/publications/future-of-compute-review/the-future-of-compute-report-of-the-review-of-independent-panel-of-experts#chap1 (2023).
92. Mytton, D. Data centre water consumption. *npj Clean Water 2021 4:1* **4**, 1–6 (2021).
93. Ludvigsen, K. G. A. The Carbon Footprint of ChatGPT. *Towards Data Science* (2022).
94. § Patterson, D. *et al.* Carbon Emissions and Large Neural Network Training. (2021).
95. Rohde, F., Gossen, M., Wagner, J. & Santarius, T. Sustainability challenges of Artificial Intelligence and Policy Implications. *Ökologisches Wirtschaften - Fachzeitschrift* **36**, 36–40 (2021).
96. Shen, K., Tong, X., Wu, T. & Zhang, F. *The next frontier for AI in China . QuantumBlack AI by McKinsey* https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-next-frontier-for-ai-in-china-could-add-600-billion-to-its-economy (2022).

97. Secretary of State for Digital, C. National AI Strategy . https://www.gov.uk/government/publications/national-ai-strategy (2021) [Accessed 8 March 2023].
98. Dawson, G. S., Desouza, K. & Denford, J. S. Understanding artificial intelligence spending by the U.S. federal government. *Brookings* https://www.brookings.edu/blog/techtank/2022/09/22/understanding-artificial-intelligence-spending-by-the-u-s-federal-government/ (2022) [Accessed 8 March 2023].
99. Boudier, J., Schmid, P. & Simon, J. Hugging Face and AWS partner to make AI more accessible. *Hugging Face* https://huggingface.co/blog/aws-partnership (2023) [Accessed 21 March 2023].
100. Mazzucato, M., Schaake, M., Krier, S. & Entsminger, J. *Governing artificial intelligence in the public interest*. https://www.ucl.ac.uk/bartlett/public-purpose/publications/working-papers (2022).
101. Meng, B. "This is China's Sputnik Moment": The Politics and Poetics of Artificial Intelligence. *https://doi.org/10.1080/1369801X.2021.2003227* (2021) doi:10.1080/1369801X.2021.2003227.
102. Zeng, J. China's Artificial Intelligence Innovation: A Top-Down National Command Approach? *Glob Policy* **12**, 399–409 (2021).
103. Webster, G., Creemers, R., Kania, E. & Triolo, P. Full Translation: China's 'New Generation Artificial Intelligence Development Plan' (2017). *Digichina* https://digichina.stanford.edu/work/full-translation-chinas-new-generation-artificial-intelligence-development-plan-2017/ (2017) [Accessed 6 March 2023].
104. Bin, R. & Dixon, L. A principled governance for emerging AI regimes: lessons from China, the European Union, and the United States. *AI and Ethics 2022* **1**, 1–18 (2022).
105. Davidson, H. 'Political propaganda': China clamps down on access to ChatGPT . *The Guardian* https://www.theguardian.com/technology/2023/feb/23/china-chatgpt-clamp-down-propaganda (2023) [Accessed 6 March 2023].
106. Shujuan, L. First domestic chatbot MOSS to be made open source. *Chinadaily.com.cn* https://global.chinadaily.com.cn/a/202302/28/WS63fd392fa31057c47ebb1211.html (2023) [Accessed 6 March 2023].
107. Marchant, G. E., Tournas, L. & Gutierrez, C. I. Governing Emerging Technologies Through Soft Law: Lessons for Artificial Intelligence. *Jurimetrics* **61**, (2020).
108. Gutierrez, C. I. & Marchant, G. E. A Global Perspective of Soft Law Programs for the Governance of Artificial Intelligence. *SSRN Electronic Journal* (2021) doi:10.2139/SSRN.3855171.
109. Pouget, H. The EU's AI Act Is Barreling Toward AI Standards That Do Not Exist - Lawfare. *Lawfare* https://www.lawfareblog.com/eus-ai-act-barreling-toward-ai-standards-do-not-exist (2023) [Accessed 2 March 2023].
110. European Commission. *AI Liability Directive*. (2022).
111. Council of the European Union. Artificial Intelligence Act: Council calls for promoting safe AI that respects fundamental rights. https://www.consilium.europa.eu/en/press/press-releases/2022/12/06/artificial-intelligence-act-council-calls-for-promoting-safe-ai-that-respects-fundamental-rights/ (2022) [Accessed 20 March 2023].
112. Government of Canada. *The Artificial Intelligence and Data Act (AIDA) – Companion document*. https://ised-isde.canada.ca/site/innovation-better-canada/en/artificial-intelligence-and-data-act-aida-companion-document (2023).
113. Kolt, N. Algorithmic Black Swans. *Washington University Law Review* **101**, (2023).
114. Cassidy, C. Australian universities to return to 'pen and paper' exams after students caught using AI to write essays . *The Guardian* (2023).
115. Jaeger, C. ChatGPT tool ban introduced at Victorian schools. *The Age* (2023).
116. Hare, J. You can't ban the bot, educators say as they struggle with ChatGPT. *Financial Review* (2023).
117. Davis, A. ChatGPT banned in WA public schools in time for start of school year. *ABC News* (2023).
118. IEEE SA. Autonomous and Intelligent Systems (AIS). *IEEE.org* https://standards.ieee.org/initiatives/autonomous-intelligence-systems/ [Accessed 8 March 2023].

119. Gebru, T. *et al.* Datasheets for datasets. *Commun ACM* **64**, 86–92 (2021).
120. Mitchell, M. *et al.* Model cards for model reporting. *FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency* 220–229 (2019) doi:10.1145/3287560.3287596.
121. OpenAI. *Best Practices for Deploying Language Models*. https://openai.com/blog/best-practices-for-deploying-language-models/ (2022).
122. Microsoft. *Foundations of assessing harm*. https://docs.microsoft.com/en-us/azure/architecture/guide/responsible-innovation/harms-modeling (2021).
123. Microsoft. *Threat Modeling AI/ML Systems and Dependencies*. https://docs.microsoft.com/en-us/security/engineering/bug-bar-aiml (2022).
124. Microsoft. *Responsible use of AI with Cognitive Services*. https://docs.microsoft.com/en-us/azure/cognitive-services/responsible-use-of-ai-overview (2022).
125. Schiffer, Z. & Newton, C. Microsoft lays off AI ethics and society team - The Verge. *Platformer* (2023).
126. International Organization for Standardization. ISO - ISO/IEC JTC 1/SC 42 - Artificial intelligence. https://www.iso.org/committee/6794475.html (2017) [Accessed 6 March 2023].
127. International Organization for Standardization. ISO - ISO/IEC 22989:2022 - Information technology — Artificial intelligence — Artificial intelligence concepts and terminology. https://www.iso.org/standard/74296.html (2022) [Accessed 6 March 2023].
128. ISO/IEC. ISO/IEC DIS 42001 - Information technology — Artificial intelligence — Management system. https://www.iso.org/standard/81230.html (2023) [Accessed 17 March 2023].
129. ISO/IEC. ISO/IEC 23894:2023 - Information technology — Artificial intelligence — Guidance on risk management. https://www.iso.org/standard/77304.html (2023) [Accessed 17 March 2023].
130. George, E. R. Corporate Social Responsibility and Social Media Corporations: Incorporating Human Rights Through Rankings, Self-Regulation and Shareholder Resolutions. *Duke Journal of Comparative & International Law, Forthcoming, University of Utah College of Law Research Paper No. 256* (2018) doi:10.2139/SSRN.3168880.
131. Ranking Digital Rights. The 2022 RDR Big Tech Scorecard. https://rankingdigitalrights.org/bts22/ (2022) [Accessed 20 March 2023].
132. Helberger, N. & Diakopoulos, N. ChatGPT and the AI Act. *Internet Policy Review* **12**, (2023).
133. Kharpal, A. China's tech giants plan ChatGPT clones — and Beijing is watching closely. *CNBC* (2023).
134. Lau, L. J., Qian, Y. & Roland, G. Reform without losers: An interpretation of China's dual-track approach to transition. *Journal of Political Economy* **108**, 120–143 (2000).
135. China Legislation Standard. Artificial Intelligence Innovation and Development Pilot Zone. http://www.cnstandards.net/index.php/artificial-intelligence-innovation-and-development-pilot-zone/ (2019) [Accessed 6 March 2023].
136. § Bello Villarino, J.-M., Gulson, K., Paul, R., Carmel, E. & Cobbe, J. Chapter X: Artificial intelligence, regulation and the purposes of education. https://www.researchgate.net/publication/368426216_Chapter_X_Artificial_intelligence_regulation_and_the_purposes_of_education_Later_version_forthcoming_in_Public_Policy_and_Artificial_Intelligence_Vantage_Points_for_Critical_Inquiry_Editors_Chapter_X_ (2023) [Accessed 3 March 2023].
137. Agence France-Presse. Google yanks gay dating app in Indonesia. *The Manila Times* (2018).
138. Reuters. Indonesia blocks Telegram messaging service over security concerns. *Reuters* (2017).
139. Reuters. Indonesia bans Chinese video app Tik Tok for 'inappropriate content'. (2018).
140. Barrett, A. M., Hendrycks, D., Newman, J. & Nonnecke, B. Actionable Guidance for High-Consequence AI Risk Management: Towards Standards Addressing AI Catastrophic Risks. (2022).
141. PDPC. Launch of AI Verify - An AI Governance Testing Framework and Toolkit. https://www.pdpc.gov.sg/news-and-events/announcements/2022/05/launch-of-ai-verify---an-ai-governance-testing-framework-and-toolkit (2022) [Accessed 17 March 2023].
142. Yeong Zee Kin. Singapore's A.I.Verify builds trust through transparency - OECD.AI. *OECD.AI Policy Observatory* https://oecd.ai/en/wonk/singapore-ai-verify (2022) [Accessed 17 March 2023].

143. Hongladarom, S. The Thailand national AI ethics guideline: an analysis. *Journal of Information, Communication and Ethics in Society* **19**, 480–491 (2021).
144. Digital Thailand. *AI Ethics Guideline*. https://www.etda.or.th/getattachment/9d370f25-f37a-4b7c-b661-48d2d730651d/Digital-Thailand-AI-Ethics-Principle-and-Guideline.pdf.aspx?lang=th-TH (2018).
145. Joshua J. Why doesn't ChatGPT know about X? *OpenAI Help Center* https://help.openai.com/en/articles/6827058-why-doesn-t-chatgpt-know-about-x (2023) [Accessed 21 March 2023].
146. Foy, P. GPT-3 Fine Tuning: Key Concepts & Use Cases. *MLQ.ai* https://www.mlq.ai/gpt-3-fine-tuning-key-concepts/ (2023) [Accessed 21 March 2023].
147. Goldfarb, A. & Trefler, D. Artificial Intelligence and International Trade. in *The Economics of Artificial Intelligence: An Agenda* vols 978-0-226-61347–5 463–492 (2019).
148. Hernandez, D. & King, R. Universities' AI Talent Poached by Tech Giants. *The Wall Street Journal* (2016).

Appendix 1: Contributing experts and peer reviewers

Lead Fellows / authors

Professor Genevieve Bell AO FTSE FAHA, Director School of Cybernetics, The Australian National University

Professor Jean Burgess FAHA, Associate Director, ARC Centre of Excellence for Automated Decision-Making and Society, Queensland University of Technology

Professor Julian Thomas FAHA, Director ARC Centre of Excellence for Automated Decision-Making and Society, RMIT University

Professor Shazia Sadiq FTSE, Director ARC Centre for Information Resilience, University of Queensland

Expert contributors

Dr Jose-Miguel Bello Y Villarino, ARC Centre of Excellence for Automated Decision-Making and Society, The University of Sydney

Dr Shakes Chandra, UQ AI Collaboratory, The University of Queensland

Dr Tong Chen, UQ AI Collaboratory, The University of Queensland

Dr Paul Dalby, Australian Institute for Machine Learning

Dr Richard Harvey, Australian Institute for Machine Learning

Professor Seth Lazar, College of Arts and Social Sciences, Australian National University

Dr Oliver Mayo FAA FTSE, Adjunct Professor, The University of Adelaide

Professor Michel Milford, Queensland University of Technology

Professor Saeid Nahavandi FTSE, Pro Vice Chancellor (Defence Technologies), Deakin University

Dr Dang Nguyen, ARC Centre of Excellence for Automated Decision Making & Society

Dr Ian Oppermann FTSE, CEO and Chief Data Scientist, NSW Data Analytics Centre

Professor Christine Parker FASSA, Melbourne Law School, The University of Melbourne

Professor Jason Potts, Blockchain Innovation Hub, RMIT University

Professor Nils Goran Arne Roos FTSE, Chair, Innovation Performance Australia Pty Ltd

Dr Aaron Snoswell, ARC Centre of Excellence for Automated Decision-Making and Society, Queensland University of Technology

Professor Nicolas Suzor, School of Law, Queensland University of Technology

James Taylor, School of Cybernetics, Australian National University

Professor Anton Van Den Hengel FTSE, Director of the Centre for Augmented Reasoning, Australian Institute for Machine Learning

Professor Kimberlee Weatherall, The University of Sydney

Distinguished Professor Mary-Anne Williams FTSE FACS, Michael J Crouch Chair for Innovation, The University of New South Wales

Associate Professor Hongzhi Yin, UQ AI Collaboratory, The University of Queensland

Professor Haiqing Yu, Professor of Media and Communication, RMIT University

Dr Junliang Yu, UQ AI Collaboratory, The University of Queensland

Associate Professor Guido Zuccon, UQ AI Collaboratory, The University of Queensland

Peer reviewers

Emeritus Professor Rod Brooks FTSE

Professor Nicole Gillespie, KPMG Chair in Organisational Trust and Professor of Management at the University of Queensland Business School

Professor Iven Mareels FTSE, Executive Dean of the Institute for Innovation, Science and Sustainability, Federation University

Dr Andrew McMullan, Chief Data and Analytics Officer, Commonwealth Bank of Australia

Professor Edward Santow, Director, Policy & Governance at the Human Technology Institute, and Industry Professor, Responsible Technology, University of Technology Sydney

Professor Toby Walsh FAA, Chief Scientist at UNSW.ai, the AI Institute, UNSW Sydney

Professor Karen Yeung, Interdisciplinary Professorial Fellow in Law, Ethics and Informatics, University of Birmingham

Conflicts of interest declaration

This briefing incorporates input from Australian experts directly involved in research in Australia. Many of these contributors and reviewers have worked directly on studies and reports cited in this briefing. Contributors and peer reviewers are drawn from a range of institutions, initiatives and fields, and collectively provide an independent and authoritative perspective on this topic.

Acknowledgements

The production of this rapid research report was supported by: Ryan Winn of the Australian Council of Learned Academies; Dr Kylie Brass and Inga Davis of the Australian Academy of the Humanities; Peter Derbyshire, Natasha Abrahams and Kylie Walker of the Australian Academy of Technology and Engineering; and Dr Hayley Teasdale, Lauren Sullivan, Chris Anderson, and Anna-Maria Arabia of the Australian Academy of Science. Edited by Robyn Diamond and Lydia Hales.

Appendix 2: Glossary

Algorithm: Automated instructions for a computer to perform a task or solve a problem.

Application program interfaces (APIs): A set of protocols to enable two software programs to communicate with one another, or for one program to run another program.

Architecture: The design or structure of an AI model.

Artificial intelligence (AI): A collection of interrelated technologies used for problem solving and to complete tasks that would otherwise require human intelligence.

AI model: A program that has been trained on a dataset to recognise patterns (typically using artificial neural networks) or reason diagnostically or predictively (as seen in probabilistic graphical models).

Artificial neural network: A type of machine learning consisting of a network of nodes that function analogously to the human brain.

Foundation models: Large AI-based models, trained on vast datasets, that can be applied to a variety of different tasks. Foundation models represent a paradigm shift in AI highlighting the phenomenal progression from algorithms (e.g. logistic regression), to architectures (e.g. transformers) to foundation models (e.g. GPT-3).

Generative AI: A type of AI model that can generate content such as text, images, audio and code, in response to user prompts.

Large language model (LLM): A type of generative AI that specialises in the generation of human-like text. May be used interchangeably with 'language models'.

Machine learning (ML): The development of models that can autonomously 'learn' from datasets and from inputs continuously.

Machine learning model operationalisation management (MLOps): A discipline concerned with the development, deployment and governance of machine learning models.

Machine learning acceleration software: Software that makes the training process for models faster.

Multimodal foundation models (MFMs): A type of generative AI that can process information from multiple types of inputs (text, visual, auditory and tactile).

Natural language processing: An interdisciplinary branch of computer science, linguistics and artificial intelligence concerned with human–computer interaction and processing using human language.

Parallel processing: Using multiple computing processors concurrently, enabling the processing of larger amounts of data in a shorter amount of time.

Technology stack: The combination of technologies used to develop an application. May be used interchangeably with 'tech stack'.

Transformer architecture: A neural network that has the feature of learning context and parallel processing, enabling more powerful and faster models. This is due to a number of underlying advances, including an encoder-decoder structure.

UX: User experience.

Vision models: A type of AI that can process visual information.

Appendix 3: Examples of regulatory actions in jurisdictions overseas

APPROACHES INVOLVING REGULATORY ACTION (LEGISLATION OR REGULATOR GUIDANCE)

Country / region	Strategy/approach	Generative AI specific implications	Implications for Australia
European Union	Proposed EU AI Act.[132] Risk management approach, categorising the applications of AI in each case to three risk categories.[h] Detailed risk management system applied to uses designated 'high risk'. Once enforced, will create legal obligation for developers and users to undertake risk management and ongoing monitoring.	Current draft text proposed by the European Council states that insofar as a system remains general purpose (i.e. without concrete outputs within a sector or task, including GPT-3 or GPT-4, or 'question-answer' systems like ChatGPT) it cannot be assessed correctly and therefore is out of scope; applications built on LLM or MFM and deployed in high-risk use cases would require compliance with risk management. Negotiations are ongoing; 'general purpose' exclusion could change before the Act is finalised. Commentators argue that LLMs/MFMs may require a 'general risk category' with appropriate obligations attached, including a general monitoring obligation for systemic risks.[113,132]	EU model has potential to become an international standard as it will apply where EU citizens are impacted by systems even if developed overseas. Also provides further impetus to international standards development: compliance with technical standards will be a way to achieve legal compliance.
China	Government-led, public-private partnership in AI development and governance, as exemplified in AI Development Plan (2017), AI Safety Framework (2020), AI code of ethics (2021), and regulatory guidelines issued on specific technologies.	The Cyberspace Administration of China has issued algorithmic and 'deep syntheses tech' regulations since 2022, all applicable to ChatGPT-style technologies.[133] China follows a 'dual-track system' to allow local experiments in policy implementation, [134] The Guidelines for National New Generation Artificial Intelligence Innovation and Development Pilot Zone Construction Work (2019) is a legal framework for companies and provinces to work on concrete aspects of AI in parallel,[135] including by: testing institutional mechanisms, policies, and regulations; promoting the in-depth integration of AI with economic and social development; and exploring new approaches to governance in the intelligent era.[136]	Reliance on guidelines and allowance for some local/provincial experimentation in governance illustrates a different approach from the EU, which is seeking to define regulatory standards applicable across the EU. China is highly engaged in developing standards for AI that have both similarities and differences from principles and approaches developed in the Global North; its approaches have potential for influence in the region.

[h] The risk levels are, unacceptable uses of AI, which are explicitly prohibited; High risk systems, which will be managed via a strong form of co-regulation, consisting of government requiring industry self-regulation via incorporation of technical standards (which do not presently exist, but are under development within the European Standards organisation, CEN-CENELEC); and a subcategory of uses subjected to transparency obligations (biometric categorisation, emotion recognition, deepfakes), which have an additional requirement for users to know they are interacting with a system.

Country / region	Strategy/approach	Generative AI specific implications	Implications for Australia
Indonesia	Applies a risk-based licensing system for all electronic service providers (Vietnam News Agency 2023). The focus of risk assessment is less on technical systems, rather on the context of use and governance. Notably, Indonesia has a history of banning apps and platforms on legal and/or moral grounds.[137–139]	The risks of generative AI may be assessed according to the nature of the content it generates at the point of licensing. Given generative AI's ability to develop unique output based on human requests, assessment is likely contingent on the perceived trade-off between economic impact of new technologies and the protection of contextually grounded moral values.	Australia should take into consideration the diverse cultural and socio-political landscape of the Asia-Pacific and consider taking a leadership role in communicating policy developments across the region.

SELF-REGULATION, VOLUNTARY, TECHNICAL STANDARDS

Country / region	Strategy/approach	Generative AI specific implications	Implications for Australia
United States	The US Government relies on self-regulation via voluntary multi-stakeholder processes for the development of risk management and technical standards. See National Institute of Standards and Technology (NIST) AI Risk Management Framework (AI RMF), v1 (January 2023). The US is also actively involved in international standards bodies.	NIST believes that its AI RMF is suitable for larger models and generative AI, treating generative systems as a category within the framework. NIST could in the future develop a specific AI RMF Profile for general purpose systems like LLMs and MFMs.[140]	The NIST AI RMF is freely and publicly available, unlike many international standards (which can be expensive to purchase), meaning it may provide useful guidance for Australian governments and firms.
Singapore	Development of standardised self-testing tools (AI Verify) under purview of the Personal Data Protection Commission (PDPC).[141,142] Aimed at enabling businesses to check the implementation of AI models against a set of principles. Singapore is also contributing to international standards development.	The framework is aimed at AI generally, rather than generative AI. The risks of generative AI might be similarly assessed by voluntary self-testing, although with qualifications.	Australia could consider encouraging businesses to conduct self-verifications of risk mitigation measures. This approach complements, rather than replaces, ethical standards in AI implementation.
Thailand	National AI Ethics Guideline provides basis for procurement-based risk management, providing principles and expectations for different actors (regulators, developers, manufacturers, end users).[143,144]	The framework is aimed at AI generally, rather than generative AI. The risks of generative AI may be assessed at the point of procurement drawing on these principles. *Ex ante* risk management operates in addition to existing laws that can regulate or deter AI misuse or harm.	Procurement-based risk assessment is an approach already employed by the jurisdiction.