Submission: Safe and responsible AI in Australia: discussion paper

To: The Minister, Australia Government Dept. of Industry, Science, and Resources.

Thank you for providing the opportunity for members of the Australian public to make submissions on this important issue. This submission is in several parts:

1. Answers to the questions in Section 5 "*How to get involved*" of the discussion paper.
2. Responses to "*Opportunities and Challenges*" in Section 3 of the discussion paper.
3. Several sections to expand on the answers to Section 5.
4. Appendices with additional information related to a "real life" example of misuse of ChatGPT that is of concern. To me this example is important because it seems to sit outside the possibility of any regulation.

My primary concern is about authorized and unauthorized use of generative AI using large language models (LLMs) by workers in safety-critical industries. I offer no comment or opinions on regulation about uses in art or creative writing, etc, even though there must be some concerns. I am sure other submissions will have addressed these concerns.

Most of my comments about LLMs can also be taken to apply to Multimodal Foundation Models (MfMs). I do not consider Automated Decision Making (ADM) at all.

References used in this submission are given in two forms.

- Web pages are referenced by links embedded directly into the text.
- Papers are listed in the Reference section at the end of the submission and are identified in the text by square brackets surrounding the number assigned to the reference [x].

Except in the examples I give, ChatGPT has not been used in the writing of this submission.

Author: John Norman

John Norman is a Western Australian independent consultant who has worked in non-destructive testing and rail flaw detection since 1979. His client base includes international companies, universities, railway inspection organizations and research organizations, as well as national companies and railway operators. He is a lifelong member of the IEEE and has served at both state and national levels in the Australian Institute for Non-Destructive Testing (AINDT). He has also been an assessor for NATA. He did early work on the use of AI in rail flaw detection in the early 1980s.

Table of Contents

Preface

In the book "*Why the West Rules – For Now*" (Profile Books, Great Britain, 2010) the author, Ian Morris, proposes the Morris Theorem: "*Change is caused by lazy, greedy, frightened people looking for easier, more profitable, and safer ways to do things. And they rarely know what they are doing*".

The introduction of generative AI into our lives is a perfect example of what Ian Morris was referring to.

Executive summary

1. National and international efforts to regulate AI are essential. However, they will not resolve risks posed by individual workers misusing generative AI tools such as ChatGPT. Workers who have used generative AI in their own time may view AI policies and regulations as being out of touch and ignore some of them. User experience with internet based AI applications may reduce trust in regulation.

2. LLM and MfM technologies have built-in limitations making them unsuitable for use in any expert system requiring accuracy, reliability, and repeatability. Such applications will need to meet high standards before being deployed. However, LLM based applications accessible on the internet will be extremely difficult to regulate.

3. Australia should contribute to international efforts to update existing standards and to develop new standards for AI applications and to regulate AI.

- Australia should follow closely the approach being developed in Europe, and if the final European Commission AI Act is acceptable, adopt its principles for an Australian act to regulate AI and encourage other, non-EU countries to do the same.
- At the international level harmonization of AI standards and regulations is essential.
- As far as is possible, Australia should update existing Australian standards, laws, and regulations in ways that are consistent with the international effort.
- Issues of jurisdiction need to be resolved
    - In my view jurisdiction should be in the country where the harm from AI occurred.

4. Australia will benefit greatly from the local development and application of AI based expert systems that do not use LLMs.

- Safety-critical industries in Australia could use AI trained for use specifically in the Australian context.
- Fewer resources are required to develop some other types of AI.
- There are opportunities for Australia in manufacturing AI enabled hardware.

5. An international approach is required to control the irresponsible deployment of generative AI applications on the internet.

Note: "Safety-critical" industries

In this submission I refer to "safety-critical" industries. I am referring to those industries where a failure would potentially be catastrophic. This includes areas like aircraft maintenance, weld inspection, electrical work, building and construction, materials testing, non-destructive testing, etc. Many industries that work to national and international standards fall under my banner of "safety-critical".

1.0 Answers to the questions in Section 5 of "*Safe and responsible AI in Australia: discussion paper*"

Definitions

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

Answer:

Yes. The definitions seem to me to be quite good and useful.

Potential gaps in approaches

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

Answer:

1. Misuse of generative AI applications available on the internet by workers in safety-critical industries: aircraft maintenance, industrial inspections, industrial laboratories, railway maintenance, etc. Worker may choose to look for answers to technical questions by using ChatGPT or similar applications. I discuss an example in the appendix.
2. Corruption of the internet by AI generated content. This has several adverse consequences.
    a. Data taken from the internet forms a substantial part of the training data used to train LLMs and MfMs. Future models will be trained on data sets containing data from previous models, which may lead to "model collapse" [10].
    b. Incorrect answers to technical questions generated by ChatGPT that get posted onto the internet come up in Google and Bing searches, making the internet a less reliable source of information for everyone. I discuss contamination of the internet in Section 2.2.4.
3. The consequences of different generative and non-generative AI systems interacting with each other. This is a risk identified by researchers at OpenAI and is discussed in the GPT-4 System Card included with the GPT-4 Technical Report [5].
4. The consequences of generative AI systems getting access to privately held data that may compromise people's privacy and safety. For example, night clubs have been collecting biometric data (fingerprints) from attendees for years. Some clubs are owned by people with criminal connections who may see advantages in (mis)using the data.
5. Australia's regulatory approach is currently not harmonized with any other country. If we end up with different countries going their own way on AI regulation, the impact of AI development will be detrimental. It will be difficult to have effective internationally agreed and accepted standards from organizations like ISO, if every country takes a different approach and trade will be affected. See Section 4.1 for more detail.

3. Are there any further non-regulatory initiatives the Australian Government could implement to support responsible AI practices in Australia? Please describe these and their benefits or impacts.

Answer:

The Australian Government must encourage work on standards, accreditation, and training, as discussed in more detail in Section 4.1.

- **Standards.** Australian industry relies on local and international standards as guides for safe and reliable work, product safety, and enabling trade. Many existing standards will need to be amended to take AI into account and there will be a need for new standards to cover aspects of AI. Australia needs to update local standards and participate in AI standards development at ISO and other standard issuing organizations. Quality management system standards like ISO9001 will need updating, for example.
- **Accreditation.** Many Australian organizations seek accreditation to ISO9001 and similar standards. Accreditation is handled by organizations like NATA and JAS ANZ. These organizations will need to factor acceptable use of AI into their accreditation processes and company audits.
- **Training.** In vocational training for any activity that works to local or international standards, students will need to be taught which AI is safe to use and which AI is unsafe because the output may be incorrect. The acceptance of AI should be conservative and cautious.

Through these processes, organizations will be able to assess whether or not particular AI applications are suitable. If a standard or accreditation requirement places restrictions on what type of AI is acceptable, responsible organizations will follow that requirement.

Also, various educational and management training courses will need to offer courses on the positives and negatives of applying AI to different enterprises. Given the known limitations of some AI, particularly LLMs, managers will need to know enough to be able to correctly evaluate all the new applications that are being developed. Managers and executives will also need to understand that they must carry some liability for making bad choices in AI selection.

4. Do you have suggestions on coordination of AI governance across government? Please outline the goals that any coordination mechanisms could achieve and how they could influence the development and uptake of AI in Australia.

Answer:

A single Federal Government department, either existing or new, should be tasked with coordination of AI deployment and regulation across the nation. AI policies need to be national policies not state policies, and where appropriate harmonized with international policies. All national, state, territory, and local governments and government owned corporations need to adopt the same approach to AI.

Differences between states will cause confusion. A uniform approach will make the development of local AI applications easier.

Responses suitable for Australia

5. Are there any governance measures being taken or considered by other countries (including any not discussed in this paper) that are relevant, adaptable and desirable for Australia?

Answer:

My view is that Australia should closely monitor what is happening in the EU, and if the European Commission AI Act [12] and the associated policies on liability [16] look suitable, Australia should adopt similar measures, and should encourage other countries to do so. Canada looks to be a good candidate as their proposed AI regulations look similar to the EU's.

I believe that harmonization is vital. The EU approach is comprehensive, except that it is weaker than the US roadmap proposed by NIST on the issue of explainable AI [18]. The final European Commission draft of the AI Act may take this into account.

The Australian Government should consider if the ideas in the document "*Blueprint for an AI Bill of Rights*" issued by the White House in the USA are appropriate for Australia. If so, an Australian "*AI Bill of Rights*" may be useful for providing legal protections to individuals adversely affected by government or private activities involving AI. See reference [21].

The Australian Government should follow New Zealand and consider the requirement for government departments and organizations thinking about introducing AI to justify the initiative. New Zealand is looking at the idea of a "social licence" for government use of AI, and justification is one part of this [23].

Target areas

6. Should different approaches apply to public and private sector use of AI technologies? If so, how should the approaches differ?

Answer:

Regulatory responses should be the same for the public and private sectors.

7. How can the Australian Government further support responsible AI practices in its own agencies?

Answer:

1. An AI Bill of Rights (reference [21]) may focus department attention on the effects of bad decisions and outcomes resulting from AI practices. We do not want another Robodebt situation. See also my answer to Question 18.
2. Do not use AI and chatbots to replace humans for people trying to contact government. It is already almost impossible to contact anyone (human) at the ATO by phone, for example. Sometimes people have issues that a chatbot cannot resolve. Just employ more people to answer phone. In this case use HI, not AI.
3. Use expert systems tailored to specific and limited applications, rather than general purpose AI systems based on LLMs and MfMs. LLMs are prone to error. LLM and MfM based AI should have restricted use within government, and not used anywhere where interaction with people is needed.

8. In what circumstances are generic solutions to the risks of AI most valuable? And in what circumstances are technology-specific solutions better? Please provide some examples.

Answer:

In general, generic solutions are adequate to control the risks of AI. A bad outcome is a bad outcome no matter what technology was used by the AI or what technology AI was incorporated into. Concerns about bias and data privacy, etc are generic. Also, concerns about unethical or illegal uses of AI are generic.

Password cracking is an example of an AI technology-neutral risk. Both ChatGPT and PassGAN can be used for password cracking, but PassGAN is optimized for password cracking and is much better at it than ChatGPT. ChatGPT is LLM based and PassGAN uses AI that is not LLM based. The use of AI for password cracking is not technology specific, but password cracking is a use that should be illegal. For more information see Section 2.2.2 and reference [6].

A technology-specific solution could apply to general purpose generative AI applications available on the internet. Because the large model based AIs are inherently unreliable, LLM and MfM AIs are unsuited for many of the applications people want to use them for. In retrospect, ChatGPT, Stable Diffusion (https://stablediffusionweb.com) and other popular generative AI applications available online should never have been released in the way they were.

An example of a technology-specific solution (other than banning these applications) might be to enforce all new generative AI applications intended for release onto the internet to initially be released online through dedicated portals designed just for that purpose. This would limit the number of people accessing the new applications and would give time to monitor the use and users to see what concerns might arise and for changes to be tested before a final release.

9. Given the importance of transparency across the AI lifecycle, please share your thoughts on:
a. where and when transparency will be most critical and valuable to mitigate potential AI risks and to improve public trust and confidence in AI?

Answer:

Transparency throughout the whole AI lifecycle is important and critical in any AI application involved with decision making. The AI lifecycle is iterative with many feedback loops, and there are many places within the cycle where the AI model is modified by training data, human input, and feedback.

Any stage where the underlying model is modified needs to be explainable in terms that can be understood by non-experts. Performance results from various stages need to be available in formats that a non-expert can understand: final tests before initial deployment, and results or feedback from early stages of deployment. In my answer to Question 8, I suggest early release of internet based AI applications through a specially designed portal that can keep the application contained and can collect data on performance. This data must be available in a form that a non-expert can understand.

The AI lifecycle is actually not a very useful concept because it has not been formalized. A better approach is to consider Technology Readiness Levels (TRLs). TRLs were developed by NASA during the Apollo program as a formalized set of development stages and have been used ever since for engineering and technical developments. A good reference was published by the Australian

Renewable Energy Agency titled "*Technology Readiness Levels for Renewable Energy Sectors*" [19]. An excellent review of the application of TRLs to AI development is given in the European Commission JRC Technical Report "*AI Watch: Revisiting Technology Readiness Levels for relevant Artificial Intelligence technologies*" [20].

TRLs spell out much more clearly than AI lifecycle diagrams the stages of a development and in the case of AI developments the points where transparency will be necessary. TRLs use nine levels with Level 9 being the final launch of the fully developed product. With AI developments, transparency requirements would apply from Level 6 (Technology Demonstration) on up.

b. mandating transparency requirements across the private and public sectors, including how these requirements could be implemented.

Answer:

Any public or private organization using AI for decision making should disclose that AI is being used with an explanation that can be accessed describing the process in simple language. If the public wish to deal with a human, to place an enquiry, or to contest a decision, the organization must have human contacts readily available to respond in a timely manner. Not like the ATO where currently it can take over 3 hours to have a phone call answered.

All AI generated content used for communication must be identified. Images should contain a watermark. Text should contain a declaration disclosing AI content. In addition, all AI generated images, text files, and audio files should have an AI acknowledgement in their metadata, similar to the way DRM is built into ebooks.

10. Do you have suggestions for:
a. Whether any high-risk AI applications or technologies should be banned completely?

Answer:

I don't see a strong case for banning certain AI technologies, but individual classes of AI products or applications could be banned.

It is too late now, but I would have no problems with ChatGPT being banned, at least temporarily, from the internet. It was released online without being sufficiently tested and has been oversold in terms of its capabilities. In fact there is a case to make that all of the current LMM and MfM AI applications on the internet should be banned until protocols are worked out that will allow their safe reintroduction.

Some deployers, for example Stability.ai (https://stability.ai) who released Stable Diffusion (a text to graphics converter) are actually opposed to regulation and their application has been used for very inappropriate purposes, including pornography. If such companies cannot act responsibly, their products should be banned.

Password breaking AI applications should be banned.

b. Criteria or requirements to identify AI applications or technologies that should be banned, and in which contexts?

Answer:

Possible criteria for banning.

- General purpose AI application with no specific end use in mind.
    - Get ChatGPT out of schools, for example.
- Being readily available on the internet for anyone to use. Some people working in safety-critical industries may ask technical questions, and the answers may be wrong.
- Known to produce errors that may be difficult for a non-expert to identify, or can be confidently wrong.
- Which use LLMs which by design do not produce reliable output.

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

Answer:

I don't think the government has any role in encouraging more people to use AI. People who need AI will find it. The government has a major role in making sure any AI is safe to use and meets appropriate standards, as outlined in earlier answers.

Regulation and accreditation will increase trust in business. It may actually decrease trust in government. Younger will probably become familiar with AI through using online applications like ChatGPT and Stable Diffusion, and through the education system if generative AI is allowed into schools. These young people with largely benign experiences with AI may view government regulation of AI as being out of touch.

Implications and infrastructure

12. How would banning high-risk activities (like social scoring or facial recognition technology in certain circumstances) impact Australia's tech sector and our trade and exports with other countries?

Answer:

Would anyone really care if Australia's tech sector were banned from developing and exporting AI applications for social scoring or facial recognition? There are many ethical applications to get involved with.

Facial recognition is actually an interesting case. There are no legitimate uses for facial recognition in everyday life, certainly not in supermarkets and hardware stores as is happening at the moment. Exporting AI facial recognition technology should not be allowed because control of the use will then be outside the reach of Australian regulation.

Australia should take the high ground on ethics when it comes to trade.

13. What changes (if any) to Australian conformity infrastructure might be required to support assurance processes to mitigate against potential AI risks?

Answer:

No specific changes are required to the conformity infrastructure. However, as explained in the answer to Question 3, there needs to be a major effort to update existing local and international standards to take AI into account, and there needs to be local and international efforts to develop the new standards required as AI becomes an everyday tool. Organizations involved with accreditation need to take AI into account as they audit companies and organizations for compliance to specific standards.

Australia should participate in this effort at every level through NATA, JAS ANZ, Standards Australia, and any other Australian organization involved with developing standards and accreditation. See Section 4.4.2 for further discussion.

Risk-based approaches

14. Do you support a risk-based approach for addressing potential AI risks? If not, is there a better approach?

Answer:

I support a risk-based approach, and I suggest that a risk-based approach will evolve by default, whether it is pursued consciously or not. Some sectors involve inherently high-risk activities. The risks can be social, psychological, technical, etc , and each class of risk may need to be handled differently. Some activities (writing poetry?) are probably intrinsically low risk.

General purpose AI applications based on LLMs and MfMs, like ChatGPT for example, can be used for both low and high-risk activities. It has been suggested in reference [13] that generative AI applications like ChatGPT , because of their special characteristics, get put into a special risk category of their own.

A risk-based approach is being pursued by the EU [12] [15] and Canada, and probably the US as well [18]. For the sake of international harmonization of regulations, Australia should also follow a risk-based approach, most likely modelled after the EU approach.

15. What do you see as the main benefits or limitations of a risk-based approach? How can any limitations be overcome?

Answer:

The main benefit is harmonization with Europe and other countries as well. This will facilitate trade and legal issues should harm from AI occur within Australia from applications developed or hosted outside of Australia. See Section 4.1 for further discussion.

I don't see many limitations, unless the majority of major trading nations cannot agree on harmonization of AI regulations and laws.

One possible limitation will be if different countries cannot agree on the hierarchy of risk. Some countries may not see general purpose generative AI applications on the internet as high-risk, but other might.

16. Is a risk-based approach better suited to some sectors, AI applications or organisations than others based on organisation size, AI maturity and resources?

Answer:

Keep it simple and apply the same approach across all sectors.

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

Answer:

Attachment C provides a good template for a risk-based approach. There are a few additional considerations though.

- **Human in the loop.** When AI technologies involving generative AI are considered, the "human in the loop" may not pick up errors or biases in the AI generated content. The whole point of the development of ChatGPT, for example, is to eventually get as human-like output as is possible. However, such output may still be "confidently wrong". It will be increasingly difficult for humans to work with such AI systems. When a problem occurs, will it be the fault of the "human in the loop", or the generative AI application that got it wrong, but convinced the human otherwise? The issue of liability will need to be addressed. For further discussion see Section 2.2.1.
- **Explainability.** As far as is practical, all AI should be explainable to some degree. The degree of explainability can depend on the risk level of the situation. Any AI used in high-risk sectors needs to be explainable for auditing purposes when a bad outcome has caused harm. Large model AI systems cannot be technically explainable because of the scale of the system, but less accurate forms of explainability may be sufficient. The document from NIST in the US *Artificial Intelligence Risk management Framework 1.0* has a good section on explainability in the context of risk management. See Section 2.2.3 and reference [17].

18. How can an AI risk-based approach be incorporated into existing assessment frameworks (like privacy) or risk management processes to streamline and reduce potential duplication?

Answer:

If I understand the question correctly, my view is that protection of privacy, as an example, must be a high priority in all dealing between citizens and government or private organizations. The risk-based approach to AI does not apply.

This will be clarified if Australia adopts an AI Bill of Rights. Such a Bill should limit how much personal information AI applications (and non-AI applications) can collect.

- On purchase and registration to use the AI product.
- In LLM and MfM training data taken from the internet.
- By limiting the ability of AI applications to bypass passwords or to override NOFOLLOW and NOINDEX metatags in websites.
- By making it illegal for professionals (e.g. doctors, lawyers, accountants, etc) to enter any client or patient private details into any online or cloud based AI application.

- By making it illegal for any government agency (e.g. police, ASIC, the ATO, etc) to enter personal or private information into any online or cloud based AI application.

Cloud-based storage is being promoted as a key building block of the rollout of generative AI applications. See reference [22] for an explanation from Microsoft.

Cloud-based storage is a huge concern for privacy. Something that has become known as Sutton's Law (https://www.fbi.gov/history/famous-cases/willie-sutton) asks "*Why to robbers rob banks?*". The Sutton answer is "*Because that is where the money is*". This law could be updated to take into account theft of data and cloud storage.

19. How might a risk-based approach apply to general purpose AI systems, such as large language models (LLMs) or multimodal foundation models (MFMs)?

Answer:

As explained in my answer to Question 14, general purpose applications involving LLMs and MfMs could be placed into a risk category of their own, especially those available on the internet. Because the risk level depends on the question asked or the task assigned to the AI, regulation will be difficult, and it may be prudent to assign a relatively high risk level to these applications, even if most use is trivial and low risk.

LLM and MfM based systems that work with other AI tools to make up expert systems with limited scope can probably be handled in the more general risk hierarchy.

20. Should a risk-based approach for responsible AI be a voluntary or self-regulation tool or be mandated through regulation? And should it apply to:
a. public or private organisations or both?
b. developers or deployers or both?

Answer:

I am not in favour of voluntary and self-regulation for anything. If a risk-based approach is applied, it should apply to everyone: public and private organizations and developers and deployers.

In general developers of new technology do not like regulation. I see voluntary and self-regulation as being useless.

2.0 Answers to the questions in Section 3 of "*Safe and responsible AI in Australia: discussion paper*": Opportunities and Challenges

The document "*Safe and responsible AI in Australia: Discussion paper*" [4] has a good section on opportunities and challenges, and a good summary of opportunities and challenges (risks) is given in "*Rapid Response Information Report: Generative AI - language models (LLMs) and multimodal foundation models (MFMs)"* [1], published by Australia's Chief Scientist.

I will briefly address some opportunities and challenges not really addressed in the above documents.

2.1 Opportunities

I see opportunities in two forms:

- Improvements to productivity through the use of expert systems that incorporate AI, and
- Benefits to the Australian economy through the local development of AI applications and experts systems, and electronic hardware that incorporate AI processors.

I believe that general purpose AI applications using LLMs and MfMs, such as ChatGPT (in any form), should not be freely available on the internet. However, there may be specialised expert systems using these technologies that can be developed to improve productivity in applications like medical image processing, engineering design, scientific investigations, etc.

There are many potential applications in medicine, mining, and technology where productivity can be greatly enhanced by use of expert systems based on AI without the use of LLMs. There are many opportunities for such expert systems to be developed in Australia and trained on Australian data.

2.1.1 Electronic hardware

I see an opportunity for Australian AI to be built into hardware for applications such as

- the Internet of Things,
- medical and scientific instruments,
- smart machine tools, and
- technology for disability support.

The current electronic hardware for such applications is generally not powerful enough to run large and complex AI software packages because in these applications smaller microcontrollers and microprocessors are used, rather than powerful computers. One way to tackle this hardware limitation is to add a specialized AI processor circuit to an existing microcontroller design, for. Some overseas semiconductor manufacturers are already doing this, see: https://www.seeedstudio.com/blog/2019/10/24/microcontrollers-for-machine-learning-and-ai/, and https://www.st.com/content/st_com/en/about/innovation---technology/artificial-intelligence.html. If Australia invests more into its electronics industry, the current interest in AI makes a perfect time to "catch the wave" of a new technological era.

As well as electronic hardware design for the AI era, Australia should be considering silicon fabrication. The Australian Strategic Policy Institute (ASPI) suggests that the time is ripe for Australia to get into silicon fabrication, see: https://www.aspistrategist.org.au/the-time-is-right-for-an-australian-semiconductor-moonshot/. High quality silicon suitable for semiconductor fabrication is

already produced and processed in Western Australia, see https://www.simcoa.com.au. The ASPI suggest establishing a local silicon foundry and **starting with simpler designs and legacy chip designs**. To start with, relatively simple microcontrollers that do not require state of the art photolithography (unlike CPUs and GPUs) could be produced locally and coupled with imported AI processors in the same package as an AI "system on a chip" or SoC. As experience and expertise is gained our designers could develop new Australian AI processor designs and fabricate them locally.

2.2 Challenges

A lot of information has been published about the challenges associated with the development of AI. The summaries in references [1] and [3] raise many challenges and risks and I am not discussing most of these in any depth.

With my "real-life" example in Appendix 1 I have identified workers misusing ChatGPT as a risk. I have not seen this discussed anywhere else and I focus on that later in this submission. In this section, I want to briefly look at just a few challenges.

- "Human in the loop".
- Irresponsible releases of AI applications onto the internet.
- Explainable AI applied to LLMs.
- Contamination of information on the internet.
- The "worker problem"
- Pressure from the large software companies to deploy AI

2.2.1 Human in the loop

Some discussions of risk with AI see a "human in the loop" as mitigation against bias and risk [1]. This is probably correct, but there are liability problems with the human in the loop approach.

When the AI employs LLMs, we know that the output can contain errors and we also know that the output can look authoritative. If the LLM based AI produces an output that is flawed in some way and the human in the loop misses it, who then is liable? It cannot be that the individual worker, the "human in the loop", is held liable for an AI generated error or bias when the fundamental basis of the tool, the LLM, is known to be authoritative and sometimes confidently wrong. **The whole point of LLM development seems to be to get to a point where the AI response is indistinguishable from a human response.**

OpenAI predicts that the problem of erroneous but authoritative output will continue with ongoing development of LLM based AI, including and beyond ChatGPT-4 [5]. This reference states "*Despite its capabilities, GPT-4 has similar limitations to earlier GPT models: it is not fully reliable (e.g. can suffer from "hallucinations"), has a limited context window, and does not learn from experience. Care should be taken when using the outputs of GPT-4, particularly in contexts where reliability is important*". It also states "*GPT-4 can also be confidently wrong in its predictions, not taking care to double-check work when it's likely to make a mistake*".

The "confidently wrong" aspect of ChatGPT and similar applications is a risk. Unless the humans in the loop can be shown to have been negligent, liability for mistakes must lay with the developers and deployers.

2.2.2 Irresponsible release of AI applications onto the internet

AI applications are being deployed without full consideration of potential harm.

- The release of ChatGPT3-5 onto the internet is an example. It was released as a "research preview" to see how users might apply it and react to it, without being fully characterized and tested. It was known that it produced errors before it was released.
- While not LLM based, PassGAN is an example of an irresponsible release of AI software by academic researchers. See https://github.com/brannondorsey/PassGAN and the associated paper [6]. PassGAN is a research project where the researchers have posted on GitHub the code for a **password guessing AI system**. With some pride, the authors note "*Further, PassGAN was selected by Dark Reading as one of the coolest hacks of 2017"*.

2.2.3 Explainability of LLMs

In safety-critical industries, when things go wrong (e.g. an aeroplane crashes or a train derails), especially if a death results, there will be at least one investigation, if not several. In such situations it is important to work out what happened in an effort to prevent it happening again. If AI is implicated, investigators will want to know how the AI application came to the answer it did. Explainable AI is required. There are ways to build explainability into AI applications. A slightly dated but good overview is given in *Explainable AI: the basics Policy briefing* [7]. A publication on explainable AI with Australian input is *Explainable Artificial Intelligence (XAI): Precepts, Methods, and Opportunities for Research in Construction* [8].

Getting developers to incorporate explainability into AI products will be a challenge, especially if it can be shown that explainability will degrade other aspects of the system performance. Further, it is believed that building explainability into ChatGPT will be very difficult because of the very large size of the LLM. As explained in *A Glimpse in ChatGPT Capabilities and its impact for AI Research* [9] "*Explainability in AI refers to the ability of an AI system to explain its decisions and provide a rationale for its output. In particular, for large language models, explainability is the ability to explain the decisions a model makes when interpreting natural language inputs. This could range from providing an explanation for a particular prediction (a posteriori) to providing a detailed analysis of the model's internal processes (a priori). The latter is simply not possible without access to the inference engine of ChatGPT and even then, it would be extremely difficult due to the large, distributed representation used in such networks. A posteriori explanation or so-called chain-of-thought (CoT) given by ChatGPT cannot really be verified nor fully trusted*".

2.2.4 Contamination of information on the internet

Research in LLMs is proceeding at a very fast rate and new concerns are arising all the time. One of these is contamination of the internet by data generated by generative AI. There are two concerns.

- A theoretical problem can occur when new AI models are trained on data containing output generated by previous versions. "Model collapse" is possible.
- Ordinary users of the internet will find even more erroneous information online than they find at the moment and trust will be reduced.

Model collapse is defined in *The Curse of Recursion: Training on Generated Data Makes Models Forget* [10]: "*Model Collapse is a degenerative process affecting generations of learned generative models, where generated data end up polluting the training set of the next generation of models; being trained on polluted data, they then mis-perceive reality*".

Content generated by ChatGPT and similar applications trained on data from the internet which is then posted on the internet will become part the data used for training future LLMs, corrupting the process. The effects of the corruption of training data has been described by some commentators in colourful language, see: https://www.downes.ca/post/75334#:~:text=As%20this%20article%20reports%2C%20%22Machine,flooded%20with%20AI-generated%20content%3F.

It will mean that in the future we may have less trust of information found on the internet and less trust in the language models that are trained on data from the internet. As an example, now when I do a Google search for an answer to the same question asked by the forum member in my "real life" example, the top response is a portion of the answer the member received from ChatGPT and posted on NDT.net, and it is wrong.

2.2.5 The "worker problem"

I am concerned that a worker in a safety-critical industry with a mobile phone can easily access general purpose generative AI applications on the internet. They can access such applications to find technical information without the knowledge of supervisors and management. I call this the "worker problem", acknowledging, however, that even a good worker under pressure may be tempted to misuse the online AI tools. This problem does not seem to be discussed or recognized elsewhere and I cannot see a high level policy approach that can prevent it.

The use of generative AI tools is unlikely to ever be accepted as compatible with quality management standards like ISO 9001. Any worker employed in a safety-critical organization with ISO 9001 certification will be going against company policy and accreditation if accessing such tools for work. However, for a variety of reasons a worker may choose to use these tools. If that happens, there will be liability issues when something goes wrong.

Deciding who to blame is not so simple.

- The individual worker?
- Management. Despite accreditation and policies, many companies have bad management which leads to workers being stressed and forced to take short cuts?
- The AI developers and deployers who have irresponsibly put generative AI tools onto the internet and promoted them for general use
    - **even though they know that the output may contain potentially dangerous errors**,
    - **who have not put in sufficient warnings and safeguards against misuse**, and
    - **who do not enforce their own usage policies**?

In my view the main liability must go to the developers and deployers, noting differences between generative AI applications where the risk of error is always going to be too high for safety-critical industries, and Google and Bing (not using Chat) searches where it is possible to access good information from reliable websites.

2.2.6 Pressure from large software companies

Sophisticated marketing campaigns, pushing AI, especially generative AI, are going to be a big problem. The large software companies see a new "gold rush" and are in a hurry to get in early with new applications. An example is Microsoft, which has invested billions of dollars into OpenAI see https://blogs.microsoft.com/blog/2023/01/23/microsoftandopenaiextendpartnership/. Microsoft

has embedded OpenAI technology into some of its own products such as Bing and Office 365. It has now embarked on a marketing campaign, including here in Australia with the recent release of "*Australia's Generative AI Opportunity*" in conjunction with the Tech Council of Australia [22].

This report is very compelling, stating that if Australia adopts generative AI quickly, it may add $115 billion to the economy by 2030, less if the pace of adoption is slower. The report ignores the caution expressed by OpenAI in reference [5] about ongoing issues with generative AI that **have yet to be resolved**. The applications Microsoft has in mind for its generative AI "copilots" seem to be exercises in deskilling the workforce. It also has in mind that it generative AI applications will store data, including personal data (e.g. patient data submitted by doctors), commercial data, and research data in its cloud based data storage system. See my answer to Question 18 for comments on privacy.

Australia needs to start getting its regulatory framework into place ASAP before people who make decisions about software purchases are seduced by the slick marketing of the big software and internet companies. Generative AI is not yet ready for such rollouts.

4.0 Regulation

The need for regulation of AI is clear, and based on the points I have made so far, more rather than less will be required. In considering regulation I am not going to specifically consider the known issues around AI in social media, population surveillance, security, education, automated decision making, education, defence, etc. I will discuss some aspects of regulation applicable to all of these areas, and at the same time I will work from my "real life" example to identify where regulation may help with the "worker problem".

I think that approaches to regulation within Australia need to take into account what is likely to happen internationally, and Australia must contribute to the international effort where it can.

4.1 Harmonization of AI regulations.

Software is an international industry, the internet is an international service, and the rollout of new AI applications will be international. It makes no sense for individual countries to have their own AI regulatory regime. Regulations and standards needs to be harmonized between countries so that developers and deployers can take a "one size fits all" approach to incorporating safeguards and quality standards into their software. Controls on AI applications that differ from country to country will be unenforceable and software for commercial AI applications will cost more.

At the moment several individual countries are looking at regulating AI, and some are already introducing local regulations and standards [4]. We learn from section 3 of the document "*Safe and Responsible use of AI in Australia: Discussion paper*" [4], that the European Union is progressing a comprehensive approach to regulation and liability [12] [15] [16].

The US is moving slowly to develop AI policies and regulations. A roadmap for the implementation of an "AI Risk Management Framework 1.0" has been established, see https://www.nist.gov/itl/ai-risk-management-framework/roadmap-nist-artificial-intelligence-risk-management-framework-ai. In my view the US approach is weak because it favours self-regulation. In Appendix D of *Artificial Intelligence Risk management Framework 1.0* [17] it states "*The AI RMF strives to:*
1. Be risk-based, resource-efficient, pro-innovation, and voluntary."
and
"7. Be outcome-focused and non-prescriptive."

As noted in Forbes online newsletter, see:
https://www.forbes.com/sites/washingtonbytes/2023/06/27/ai-regulation-is-coming-to-the-us-albeit-slowly/?sh=6aec2caa7ee1, "*By moving slowly, U.S. firms are often at the mercy of EU regulations, which often come into effect more quickly and become the de facto industry standard. For example, research shows that after the EU passed the General Data Protection Regulation (GDPR), even those companies not covered by the GDPR's rules, such as those based in the U.S. with customers also based in the U.S., changed the way that they collect and store personal data*". This is not a bad thing. If US companies are forced by circumstances to follow the EU lead, then countries planning AI regulations, including Australia, should do the same for the sake of harmonization.

The NIST roadmap document on risk management [17] explains that explainability is important, noting "*Risk from lack of explainability may be managed by describing how AI systems function, with descriptions tailored to individual differences such as the user's role, knowledge, and skill level. Explainable systems can be debugged and monitored more easily, and they lend themselves to more thorough documentation, audit, and governance.*".

Reference [17] also notes "*Transparency, explainability, and interpretability are distinct characteristics that support each other. Transparency can answer the question of "what happened" in the system. Explainability can answer the question of "how" a decision was made in the system. Interpretability can answer the question of "why" a decision was made by the system and its meaning or context to the user*".

The EU documentation associated with the AI Act seems to place very little emphasis on explainability of AI and they should consider the US approach.

The European Commission has given a lot of attention to risk. In its proposed AI Act [12], the European Commission has established a hierarchy of risk: minimal, limited, high, and unacceptable. Applications with unacceptable risk will be banned. Applications with high risk will be managed through regulation.

In [16] the European Commission defines the higher level risks as follows.

"*Unacceptable risk*
All AI systems considered a clear threat to the safety, livelihoods and rights of people will be banned, from social scoring by governments to toys using voice assistance that encourages dangerous behaviour.

High risk
AI systems identified as high-risk include AI technology used in:

- *critical infrastructures (e.g. transport), that could put the life and health of citizens at risk;*
- *educational or vocational training, that may determine the access to education and professional course of someone's life (e.g. scoring of exams);*
- *safety components of products (e.g. AI application in robot-assisted surgery);*
- *employment, management of workers and access to self-employment (e.g. CV-sorting software for recruitment procedures);*
- *essential private and public services (e.g. credit scoring denying citizens opportunity to obtain a loan);*
- *law enforcement that may interfere with people's fundamental rights (e.g. evaluation of their liability of evidence);*
- *migration, asylum and border control management (e.g. verification of authenticity of travel documents);*
- *administration of justice and democratic processes (e.g. applying the law to a concrete set of facts).*

High-risk AI systems will be subject to strict obligations before they can be put on the market:

- *adequate risk assessment and mitigation systems;*
- *high quality of the datasets feeding the system to minimise risks and discriminatory outcomes;*
- *logging of activity to ensure traceability of results;*
- *detailed documentation providing all information necessary on the system and its purpose for authorities to assess its compliance;*
- *clear and adequate information to the user;*
- *appropriate human oversight measures to minimise risk;*
- *high level of robustness, security and accuracy.*"

The points "*critical infrastructures (e.g. transport), that could put the life and health of citizens at risk"* and "*safety components of products"* could do with some expansion to cover more clearly the safety-critical industries that I am concerned about. There is also no specific mention of generative AI, but the second part on "High-risk AI systems" will exclude applications like ChatGPT, if it can be classed high-risk.

In reference [14] the author argues that generative AI should be placed in a risk category of its own: "*For this reason, rather than trying to fit general-purpose AI systems into existing high-risk categories, we propose that they should be* ***considered a general-risk category in their own right,*** *similar to the way that chatbots and deep fakes are considered a separate risk category of their own, and subject to legal obligations and requirements that fit their characteristics.*"

This is sensible.

The European Commission is considering the issue of liability [16] in relation to AI. This, to me, is important. As discussed in [16], it will not always be easy to identify who is at fault when an AI system is implicated in a bad outcome. My belief is that developers and deployers of generative AI applications carry liability in the countries where they are used and must be subject to local laws. People harmed by generative AI must be able to use their national court system to make a claim for damages and compensation, not in the jurisdiction nominated by the developer or deployer of the AI application in their terms of use. This seems to be the approach taken in the European Commission proposal.

OpenAI has lobbied in Europe against general purpose AI based on large models being classified as high-risk [14], but they admit that ChatGPT is capable of being used for high-risk activities. The European Commission has watered down some of their proposed regulations in response to lobbying by OpenAI.

The development of EU regulations for AI is still a work in progress. I believe that Australia should follow their progress. If the end result looks reasonable, Australia should adopt the same regulations, as should every other country. In my view, harmonization of regulations is essential.

4.2 An Australian approach to regulation

4.2.1 Establish a Federal Government body specifically to regulate AI in Australia.

This body would regulate AI in Australia. It would also handle complaints about possible misuse of AI and would investigate incidents where AI has been implicated in bad outcomes.

Over time, such a regulator may be able to deal with the bad worker in safety-critical industries problem. First, it needs to be established through investigations of AI related incidents that it is a real problem. If it is a real problem then where appropriate the law needs to be applied, even to generative AI deployers and developers who may be located in other countries. In my view, in most such incidents the least liable will be the actual worker, who, despite company policy and national regulations, will be encouraged by the promoters of generative AI to use the services available on the internet. Younger workers may actually come to view company policies and government regulations concerning the use of generative AI as being out of touch.

4.2.2 Participate in international efforts to regulate AI

Australia should participate where it can in international bodies that will be set up to monitor and regulate AI. Australians should be involved in developing international standards involved with AI. The International Organization for Standardization (ISO) has a Joint Technical Committee working on standards for AI, see https://www.iso.org/committee/6794475.html. Australia is a member of ISO should be involved at every level. Australia should also be involved with the AI ethics work at UNESCO, see https://www.unesco.org/en/artificial-intelligence, for example.

None of this good work will directly affect the worker problem, but Australia must be involved at every level of developing regulatory policies to help control the development and rollout of AI applications to mitigate all the other problems that have become apparent. Australia must do what it can to ensure that current and future general purpose generative AI applications on the internet are classified as intrinsically high risk. Australia must work with other nations to help develop tools that can evaluate the risk level of generative AI applications.

4.2.3 Place some restrictions on general purpose generative AI applications available on the internet

These may be difficult to achieve as a single nation, but Australia should consider the following in its international efforts.

Any general purpose generative AI application, like ChatGPT, must have the following.

- Built-in blocks to prevent certain high risk outputs, e.g. aircraft maintenance. There are some blocks in ChatGPT at the moment, but I have seen nothing in my experiments that prevented me getting incorrect answers to technical questions of the type a worker in a safety-critical industry might ask.
- Warnings of potential high risk for other outputs that are associated with technical issues from safety-critical industries.
- Referencing. All general purpose generative AI applications on the internet must give references and citations.
    - The OpenAI release of ChatGPT does not give references and citations. In a sense, it cannot give specific references because, if I understand correctly, the OpenAI LLM does not associate word collocations with specific sources. OpenAI must find a way around this.
    - The Bing Chat function does provide references and citations because Microsoft have incorporated ChatGPT into a Microsoft AI application called Prometheus which allows this, see https://searchengineland.com/microsoft-bing-explains-how-bing-ai-chat-leverages-chatgpt-and-bing-search-with-prometheus-393437#:~:text=Citations%20and%20links.,because%20of%20the%20Prometheus%20technology.
- Restrictions must be placed on the data used for training future internet based LLM applications. As these applications become more popular people will post output onto other websites, and tools may even be developed specifically to post AI generated information into online forums, newsfeeds, blogs, etc. This AI generated content will then be used for training future LLMs. Much of the AI generated content will contain errors which will be perpetuated. The internet is currently a good source of information for people who know how to find and evaluate it. However, there is a lot of incorrect information available and this will become more of a problem as AI generated content becomes a significant part of what is available. The idea of "model collapse" has already been discussed in an earlier section of this submission.

- The general purpose generative AI applications on the internet must be explainable in some way. In [17] it states "*Explainability refers to a representation of the mechanisms underlying AI systems' operation,*". There are various levels of "explainability" some of which involve getting a picture of how the learning networks are configured by the training to produce the outputs. This technical approach is seen to be impractical for LLMs because of the sheer size of the model. A useful guide to the goals of explainability is given in [18] where the approach is less technically rigorous and may be more applicable to LLMs and MfMs: "*Four Principles of Explainable Artificial Intelligence".* The authors explain "*We introduce four principles for explainable artificial intelligence (AI) that comprise fundamental properties for explainable AI systems. We propose that explainable AI systems deliver accompanying evidence or reasons for outcomes and processes; provide explanations that are understandable to individual users; provide explanations that correctly reflect the system's process for generating the output; and that a system only operates under conditions for which it was designed and when it reaches sufficient confidence in its output*".

4.2.4 Insurance and jurisdiction

In [13] the authors write "*We argue that generative AI systems such as ChatGPT differ on at least two important points from the 'traditional' AI systems the Act has originally been written for: dynamic context and scale of use. Generative AI systems are not built for a specific context or conditions of use, and their openness and ease of control allow for unprecedented scale of use. The output of generative AI systems can be interpreted as media (text, audio, and video) by people with ordinary communication skills, lowering, therefore, significantly the threshold of who can be a user. And they can be used for such a variety of reasons to some extent because of the sheer scale of extraction of data that went into their training*".

Further, they write "*The whole point about generative AI as a general-purpose AI system is that because they can be used for so many different purposes, it is paramount to incentivise the providers of systems to think about the safety of these systems from the onset, starting with the difficult question of data quality*".

One way that the European Commission is planning to "incentivise" providers of systems is to update the EU civil liability laws to take into account the special features of AI [11] [16]. The proposal is that, depending on the risk of the system or application, the amended rules will allow a simpler path to seek compensation for damage arising as a result of failure of the AI or misapplication in relation to EU rules. In higher risk applications, to developers or deployers will be required to disclose certain information in a way that takes the burden off the person laying the complaint. They are also considering mandatory insurance. The EU AI rules will over-ride any terms of use set by the providers like OpenAI and Microsoft and will force them to deploy carefully in Europe.

The Australian government must also require deployers of general purpose generative AI products, either released in Australia or accessible from Australia via the internet, to carry mandatory professional indemnity insurance (they "give" advice) and public liability insurance (they produce errors) in Australia.

Australia regulations must also over-ride the statements of jurisdiction that appear in the terms of use from providers. The over-riding of jurisdiction is proposed in [16] for the EU and Australia must do the same and harmonize Australia's regulations with those of the EU.

Appendix 1. Example of misuse of ChatGPT

This is an example of the misuse of ChatGPT by an employee working in Non-Destructive Testing (NDT) an industry that is involved with safety. My professional experience is in NDT.

A1.1 Context

NDT involves ways to inspect materials, components, structures, and systems for defects without causing damage. It uses various technologies such as ultrasound and radiography that are non-destructive. The object is to identify manufacturing or service flaws and defects in anything ranging from printed circuit boards to bridges, aircraft, railway materials, submarines, pipelines, wind turbine blades and structures, mining machinery, storage tanks, and used everywhere from power stations to oil rigs, nuclear reactors, etc. A major aspect of NDT involves inspecting welds and another involves looking for loss of material due to corrosion.

As in all safety-critical industries, NDT inspectors must have certification based on training, education, and experience and this is defined in national and international standards.
Many industries and organizations are required to work to very high technical standards and if these high standards are not met safety can be compromised, potentially with disastrous results. Aircraft maintenance is an example. In this submission I call these industries "safety-critical" industries. The safety-critical work may be performed in-house or by contractors. For this submission, I take safety-critical industries to be those that include activities such as laboratory testing, calibration, inspection, product certification, maintenance, electrical installation, etc. Many such organizations seek accreditation from NATA, JAS ANZ, and similar bodies to provide documentary evidence that they meet the standards required for their line of work. They work to national and international standards and often use detailed technical procedures provided by the manufacturers of the equipment they use or the systems they work on.

My concern with the following example is that I cannot see a way for regulation to prevent this risk. I don't see any of the regulatory approaches raised in "*Safe and responsible AI in Australia: Discussion paper*" [4] as being useful for this problem.

A1.2 The example

Many NDT practitioners follow an online forum run by "NDT.net" (https://www.ndt.net/forum/forum.php), an organization based in Germany. Practitioners can ask NDT related questions and experts (even myself, occasionally) will try to answer them. The following concerns a recent thread on the forum.

On April 17, 2023 a member asked the forum for an explanation: "*for what's the focal distance of a TR probe?*" and also "*how can we measure this focal distance in practice?"*.

Following the thread it seems that the member had initially asked ChatGPT for a way to determine the focal length of a TR probe (an ultrasonic Transmit/Receive transducer used in NDT inspections for measuring the thickness of materials or finding corrosion) and did not understand the answer. She decided to ask the forum how to do it.

The technical details of the answer from ChatGPT are not important except to note that the ChatGPT answer refers to “the knife edge method” which is wrong. The “knife edge” method is generally used in optics and cannot be used to find the focal length of an ultrasonic TR probe. However, the ChatGPT answer looks authoritative and is written in a style similar to an NDT procedure.

A1.3 What are the problems?

1. The member should have known how to determine the focal distance of a TR probe. It is part of basic training in ultrasonic NDT.
2. The ChatGPT answer was wrong but expressed in a confident, authoritative style.
3. NDT is a high-risk activity because getting it wrong in some way can lead to serious harm: aircraft crashes, train derailments, pipe failures in nuclear reactors (think 3 Mile Island), etc. ChatGPT should have either blocked the question or flagged warnings, in accord with the OpenAI Usage Policies.
4. An indirect problem is that the output from ChatGPT that the member posted on NDT.net now comes up at the top of a Google search for the same question the member asked. The incorrect information supplied by ChatGPT is now part of the body of knowledge available on the internet. See Section 2.2.4 which discusses contamination of the internet for more on why this is a problem.

A1.3.1 The worker problem

We don’t know much about the member and the reason for asking ChatGPT the question about TR probes, but it looks like the member wanted a genuine answer. The member’s profile on NDT.net, shows supposed proficiency in two main NDT methods: Acoustic Emission (AE) and Ultrasonic Testing (UT). AE is a sophisticated condition monitoring technique (detecting shockwaves from growing cracks using special ultrasonic microphones) that requires considerable training, so it is clear that the member is no fool.

It is lucky that the member found the ChatGPT answer confusing and asked the forum. In the circumstances this was the right thing to do. However, if the ChatGPT answer had been a bit less confusing but still incorrect, would the member have botched the calibration before carrying out an NDT inspection? Would the inspection have missed something critical which later led to a failure, perhaps in a pipeline or an aircraft component?

Workers will looks for answers if they are given problems they do not know the answer to. In safety-critical industries, where those answers come from and the reliability of the answers is important. Not every worker is motivated to do well and not every worker has adsorbed all the lessons from their training. Sometimes workers are given tasks that they are not qualified to do or are put under pressure to do without sufficient time or resources.

It is a problem when workers in safety-critical industries seek answers to technical questions from unreliable sources, including ChatGPT, despite organizational policies and regulations that forbid it.

A1.3.2 The output from ChatGPT

ChatGPT produces output with appropriate tone and style for the context of the question asked and it uses technical words and phrases to add detail. It cannot analyse technical detail for accuracy and

appropriateness. Nor can it analyse the logic of the sequence of steps in the procedure it produces. The output can be "confidently wrong" [5].

ChatGPT output looks authoritative, but without further checking is unsafe to use. A worker who is poorly trained or tired may be unable to identify the errors and will use the output "as is". It is not a technical resource that people working in safety-critical industries should ever use to get reliable information. It is not designed to do this [1].

Safety-critical industries rely on detailed technical information specific to particular applications and systems and which is often proprietary with controlled distribution, and they also rely on national and international standards. These resources can never be replaced by an LLM based AI system trained wholly or partly on internet data.

The training data for the LLM of ChatGPT, some taken from the internet by Common Crawl [3], does not cover the kind of technical material needed for safe answers to NDT related questions. Even if trained on ideal material, the LLM produces natural language output by predicting what word or punctuation will come next, using the prompt to identify the context, and will never reliably relay technical information to the user to a level suitable for safety-critical industries.

A1.3.3 OpenAI/ChatGPT Usage Guidelines

OpenAI has published Usage Policies, see https://openai.com/policies/usage-policies. These have been updated regularly since the initial ChatGPT release. The latest version is dated 23/03/2023.

The Usage Policies include a prohibition on "**Activity that has a high risk of physical harm: including**….**Management or operation of critical infrastructure in energy, transportation, and water**". NDT is used extensively in all of these sectors and should be considered high risk by default.

The current online version of ChatGPT will recognize some NDT related questions (I have asked it many) and suggest using an experienced NDT technician to carry out the procedure, but nothing has been flagged high risk or blocked. In general, the usage policies are not enforced in any meaningful way.

GPT-4 is meant to be better at identifying risk that the free online version of ChatGPT [5]. It should refuse to answer many questions that it identifies as going against usage policy. However, users are able to get around many of these restrictions using a process known as "jailbreaking", see https://www.techopedia.com/what-is-jailbreaking-in-ai-models-like-chatgpt. Jailbreaks are known to be a problem in GPT-4 [5].

References

[1] 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.

[2] Hajkowicz SA1+, Karimi S1, Wark T1, Chen C1, Evans M1, Rens N3, Dawson D1, Charlton A2, Brennan T2, Moffatt C2, Srikumar S2, Tong KJ2 (2019). *Artificial intelligence: Solving problems, growing the economy and improving our quality of life*. CSIRO Data61, Australia.

[3] Zhou,W, et al (2023, May 7). *A Survey of Large Language Models*. arXiv: 2303.18223v10.

[4] Australian Government, Department of Industry, Science, and Resources (June 2023). *Safe and responsible AI in Australia: Discussion paper*. https://consult.industry.gov.au/supporting-responsible-ai.

[5] OpenAI (2023, March 27). *GPT-4 Technical Report*. arXiv:2303.08774v3.From https://openai.com/research.

[6] Hitaj, B, et al (2019, February 14). *PassGAN: A Deep Learning Approach for Password Guessing*. arXiv: 1709.00440v3.

[7] The Royal Society (2019, November). *Explainable AI: the basics Policy briefing*. ISBN: 978-1-78252-433-5

[8] Love, P. et al (2022). *Explainable Artificial Intelligence (XAI): Precepts, Methods, and Opportunities for Research in Construction*. arXiv: 2211.06579

[9] Joublin,F. et al (2023, May 10). *A Glimpse in ChatGPT Capabilities and its impact for AI Research*. arXiv: 2305.06087

[10] Shumailov, I. et al (2023, May 31). *The Curse of Recursion: Training on Generated Data Makes Models Forget*. arXiv: 2305.17493v2

[11] Madliega, T. (2023, Feb). *Briefing: Artificial intelligence liability directive*. European Parliamentary Research Service PE739.342

[12] European Commission (2021, April 21). *Proposal for a Regulation of the European Parliament and of the Council Laying down harmonized rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union legislative acts*. Brussels. COM(2021) 206 final. 2021/0106 (COD).

[13] Helberger, N. (2023, February 16). *ChatGPT and the AI Act*. Internet Policy Review, Vol. 12(1) https://policyreview.info/essay/chatgpt-and-ai-act.

[14] OpenAI (2022) *OpenAI White Paper on the European Union's Artificial Intelligence Act*. ares 20226851313 (unpublished)

[15] European Commission (2013, June 20). *Regulatory framework proposal on artificialintelligence. Shaping Europe's digital future*. https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai.

[16] European Commission (2022, September 28). *Proposal for a Directive of the European Parliament and of the Council on adapting non-contractual civil liability rules to artificial intelligence* . Brussels. COM(2022) 496 final. 2022/0303 (COD).

[17] US Department of Commerce, National Institute of Standards and Technology (2023, January). *Artificial Intelligence Risk Management Framework 1.0*. https://doi.org/10.6028/NIST.AI.100-1.

[18] US Department of Commerce, National Institute of Standards and Technology (2021, September). *Four Principles of Explainable Artificial Intelligence*. https://doi.org/10.6028/NIST.IR.8312.

[19] Australian Renewable Energy Agency (2014, February). "*Technology Readiness Levels for Renewable Energy Sectors*".

[20] European Commission JRC Technical Report (2022) "*AI Watch: Revisiting Technology Readiness Levels for relevant Artificial Intelligence technologies*". JRC129399. EUR 31066 EN. https://ec.europa.eu/jrc.

[21] The White House (2022, October). "*Blueprint for an AI Bill of Rights*". https://www.whitehouse.gov/ostp/ai-bill-of-rights/.

[22] Microsoft and Tech Council of Australia. (2023, July). "Australia's Generative AI opportunity". https://techcouncil.com.au/newsroom/generative-ai-could-contribute-115-billion-annually-to-australias-economy-by-2030/.

[23] World Economic Forum (2020, June). "*Reimagining Regulation for the Age of AI: New Zealand pilot project*". White paper.