Submission
Safe and responsible AI in Australia

Submission by Iris Vardi

This submission contains responses to selected consultation questions. These responses have been in a large part informed by my attached case study of the roll out of ChatGPT. The main findings from my case study are summarised in the executive summary. Further detail can be found in the case study itself.

Declaration

This document is entirely the work of the author without any generated language from any LLM or any other form of AI.

1 Contents

2 Responses to Selected Consultation Questions

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

1. Ensure that trust is not excessive through thorough education on the abilities and limitations of AI technology that is being used.

2. Ensure that people at any level of the organisation can challenge an AI output or decision. For example:

   - Ensure a high knowledge base and skills about the job to the level that it could be done without AI;

   - Ensure processes and a culture for challenging decisions.

3. When generative AI is released worldwide as occurred with ChatGPT, and public servants find aspects of that technology useful, then an appropriate version of that technology should be customised for internal use only so that employees do not use the technology on the internet. That approach should also apply for businesses.

4. As a matter of course on each and every document and website, both in business and in government, a notice should be included which details whether or not AI including generative AI was used and to what extent, and the role of human oversight in the document for the purposes of being able to judge and evaluate the document. For instance:

   This document was in part generated by <name of LLM>, a language generation model. Data was gathered and analysed by <name of AI system>, an AI system which … This analysis formed part of the input into <name of LLM>. The AI data, its analysis and the AI generated language was reviewed by ….

   This document includes text generated by <name of LLM>, a language generation model and therefore the reader should check the veracity of the content. This content cannot be used for legal purposes …

   This document is entirely the work of the author without any generated language from any LLM or any other form of AI.

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

1. Ensure a strong legislative framework with strong regulation that applies through the entire AI cycle right from the initial research release phase (see risks that arose from research release in the attached case study), through to approval mechanisms if required, through to the actions by the deployers, and then finally to the organisation using it.

   By applying a strong legislative framework, both the AI development companies and

businesses using these products can operate in an equal playing field providing commercial certainty and confidence to all.

2. Trust in a comprehensive regulatory approach from the initial research releases through the entire AI cycle will be enhanced if each and every enterprise, business, and government department:

   - has a culture supported by processes that enables individuals both within the organisation and external to the organisation the ability to challenge any and all AI/LLM outputs in an easy straight forward way;

   - has clear easy mechanisms for clients or recipients of AI or LLM generated outputs and decisions to speak to humans who have the knowledge, skills and authority to investigate and address issues.

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

1. This is a good approach which needs further development. As presented in Box 4 of the discussion paper, however, it is very reactive, looking at risks only after the AI technology has already been deployed in relation to the purpose for which it is being used. This could leave us on the backfoot, sometimes making it too late to put proper controls in place. Therefore, this risk-based approach should apply to the whole AI cycle.

2. A proactive stance in relation to risk would improve this approach greatly. This would be enabled by applying a risk-based approach to whole AI cycle. One important proactive stance would be looking at risks arising right from the start of the development of an AI product, and hence the regulatory controls that need to be put in place both globally and federally. For instance,

   - Regulation when research to improve an AI application involves members of the public anyone from outside the AI company. This is particularly important in developing generative AI so that its deployment during the research phase does not by default become a social experiment with serious irreversible consequences. For example:
     - Ensuring that human research protocols are followed such as informed consent
     - Getting ethics approval

     See the attached case study with details of how OpenAI's ChatGPT research preview release to the public in an uncontrolled manner has led to a wide range of serious consequences, some already irreversible, along with the need for regulatory control in this area.

   - Regulatory requirements that need to be met prior to release of an AI product particularly generative AI. For certain types of AI this may include going through an approval process for safety such as occurs prior to the release of medications on to the market. This may include some of the applications listed in the medium risk category in Box 4 of the consultation document.

- Regulatory requirements that need to be met prior to AI being deployed within an organisation such as
    - Ensuring that generative AI is physically contained within an organisation or an industry group
    - Ensuring that generative AI systems are unable to communicate with another AI system to mitigate against the possibility of one AI system training the other

3. This risk approach should also be augmented by an approach that looks at wider societal and business protections, addressing issues such as:

- Strategies and approaches to mitigating the deskilling of the population so that human oversight, judgement and evaluation of AI outputs can be effectively managed;
- Protecting our knowledge base and ensuring free and open access to sources to be able to judge, evaluate and challenge AI decisions. See risks to knowledge base addressed in the attached case study;
- Strategies for overcoming deleterious impacts on relationships and trust between the public and business/government organisations including impacts of automation bias. See the risks in the attached case study;
- Strategies for ensuring that the public is fully aware that they are dealing with a generative AI system, in particular LLMs;
- Significant penalties for anyone who trains systems to act in what could be perceived to be deceitful manner;
- Legislation to ensure that mechanisms exist to disable any form of AI should that become necessary – this is extremely important to the build of any AI system.

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

1. Adding a proactive approach and a societal protection approach to the proposed risk-based model will help to address general purpose AI systems throughout the AI lifecycle. Regulation needs to be world-wide due to the serious risks that have already arisen from the uncontrolled release of ChatGPT. These worldwide generative AI products should be considered in the high-risk category. See attached case study.

   This should start from the iterative deployment phase when the product is essentially still in development and undergoing further research. Once the technology is ready for feedback from and interactions with humans outside of the organisation then its deployment also becomes a social experiment. At this point legislation and appropriate regulatory controls need to be applied with the developers required to adhere the same ethical standards as required in other research settings such as universities, hospitals and the like.

This includes the requirement for:

- Research with humans to be conducted in a controlled way, employing all the proper steps for recruitment to a research project, including controlling size of the trial, call for volunteers, full disclosure of how the system works and the role and responsibilities of the volunteers, along with signed informed consent and so forth

- AI research with humans to only be conducted with people old enough to give consent, i.e. 18 or over.

2. A defensive stance to risks needs to be taken with regards worldwide generative AI particularly as its abilities will improve over time and its reach will become wider. For instance, serious consideration should be given to the following:
    - How much data is held, where it is held and for how long

    - Reconsidering cloud storage which may not be safe at all as generative AI may well learn to break passwords and access data without gaining any permissions to do so and there is no way to punish a generative AI system or put sanctions on it

    - Regulatory controls to physically contain generative AI systems and keep them separate so that they do not communicate with each other and thus train each other;

    - Regulatory controls to ensure that knowledge is protected in the deployment of generative AI, for instance, banning the use of AI generated text to be displayed at the top of search engines due to the implication that the generated text was actually information obtained in a search. See issues with implicit limitations in the attached case study showing the confusion between search and AI generated text.

    - Regulatory controls on AI developers requiring them to provide education modules for each and every user prior to them using the system that clearly and simply goes through how the system works, what it can do and what it cannot do and what it must not be used for. See the accompanying case study which highlights how plain language notices and 'fine print' do not work and how misunderstanding the technology can lead to significant real-world problems

    - Mitigation against AI companies using their Terms and Conditions as a means of dealing with misuses of a generative AI product which results in significant real-world problems rather than taking responsibility. See the attached case study showing how real-world problems were dealt with through updating the Terms and Conditions of Use.

2.5 Question 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?

1. A broadened risk-based approach that applies to the whole AI cycle, which includes both reactive and proactive stances along with protections, should be mandated through legislation

supported by strong regulatory controls. It should apply to public and private organisations, developers and deployers. Risks will already be reduced substantially if regulation is applied to developers and deployers so that a safer product is deployed in a contained safe manner to begin with.

Regulation must be mandated. Despite the multiple problems seen with the premature release of ChatGPT (see attached case study), Microsoft, the main financial backer of OpenAI, is rushing out prematurely promoting and releasing LLMs in Bing, in schools and the like. While more than 1100 signatories, many with significant technical knowledge in the area, signed an open letter to pause AI development (Loizos, 2023), Microsoft dismissed most if not all of their entire ethics team (Bellan, 2023). Therefore, it is incumbent on us to provide the mandated regulation and oversight that some of these AI companies are unwilling to do.

2. Everyone needs to be required to take responsibility and to be accountable for their actions in relation to the research, the release, the deployment and the use of AI. Doing so will provide an even playing field for all. Voluntary codes and self-regulation do not provide an even playing field as commercial interests and commercial advantage will over ride these. We are already seeing this in the roll out of ChatGPT long before it is ready and before protections have been put in place leaving OpenAI developers, business, government and education on the backfoot. OpenAI developers are rushing to update notices and systems 'on the run' as business, government and education are rushing to deal with the fallout; reacting not controlling.

3 Case Study
Executive Summary

This case study looks at the risks and lessons learnt from the way in which people engaged with OpenAI's premature worldwide research release of their ChatGPT and the actions undertaken by the company.

This examination reveals that:

- Many people did not understand the intent of the release and how the system worked, and hence used it for purposes which the LLM was unable to reliably do;
- This led to breaches of confidentiality, falsehoods being spread and decisions being made on the basis of faulty information with significant real-world risks;
- Plain language notices and terms and conditions of use were largely ignored;
- The LLM was used by a significant number of people in ways that were not intended and for which the program had not been designed;
- In large part the problems arose due to the uncontrolled nature of the release during its research phase resulting by default in a large social experiment;
- Excessive trust in the LLM by naïve Users appeared to be due to not understanding the stage of the AI lifecycle that the product was in, the human like responses received and the impressiveness and coherence of the output which was then not checked;
- Many risks were revealed including risks to knowledge base, risks from incorrect use, and risks from blind acceptance due to excessive trust.

On the basis of this examination, the following is suggested:

- that the wide release of generative AI in the research phase is banned;
- that any AI research release involving human interaction outside of the AI company is subject to the same ethical standards and requirements of human research including ethical approval, informed consent and appropriate training before the Users themselves either directly or indirectly train the generative AI;
- that iterative deployment of generative AI with the population at large needs also to be considered a human experiment and as such needs to be controlled in the same way that we conduct research involving humans including ethical approval, controlled release, informed consent and so forth;
- that AI products are subject to thorough regulatory controls with prescribed standards prior to any release;
- that Users are educated on each and every LLM product prior to its use to understand the limitations and why they exist, and to empower individuals to question outputs;

- that it becomes general practice that each and every document and website include a statement declaring whether or not generative AI or an LLM was used in its production and the extent to which it was used;

- that actions are taken to protect not only our information, but also our sources and our knowledge base, and ensure free access to that knowledge and information;

- that clear parameters are provided for the actions that AI companies need to take to protect society at large and Users, including thorough educational programmes (i) to mitigate against misuse and (ii) to address misuse when it occurs;

- that practices within organisations ensure the retention of human control along with the skills, knowledge and ability to challenge;

- that the public can always speak to a human who is skilled in the matters to do with the organisation, and who is empowered to make decisions and to investigate situations arising;

- that we continue to develop and retain skills and knowledge that we currently have in order to be able to put in that capable human oversight including the ability and skills to gather and search for information, organise information, analyse information, make judgements and predictions within a given discipline area, field of research, profession or job of any type without needing to rely on AI.

4 Case Study Introduction

This submission centres on human behaviour primarily in relation to generative AI, particularly Large Language Models (LLMs) such as ChatGPT by the companies that release these products and on the subsequent inappropriate and incorrect use of Generative AI by everyday people in their private, work and study lives. While there are and there will be useful applications for AI, this submission focusses on the significant risks arising from human behaviours and their implications for practices, regulation and legislation at the personal, business and governmental levels to ensure safe and responsible AI.

In this submission, I will be focussing on the release of ChatGPT to the public by OpenAI, but only as an example. Clearly there are and there will be other LLMs, both written and spoken, developed by other companies, which mimic human conversation, writing and speech. This release provides some lessons and insights that are worth examining with a view to determining some of the risks and controls that need to be put in place.

On November 30 2022, OpenAI released, free to the public, a research preview of ChatGPT. In conducting their LLM research publicly, however, this release has also become a large unregulated social experiment which is already leading to real world problems as a significant number of everyday Users and commentators misunderstand the intent of this release of ChatGPT, the data on which it relies and its capabilities in the real world.

It is interesting to note that OpenAI, on the whole, has understood many of ChatGPT's serious limitations and has placed significant caveats on its use. Some of these limitations and caveats are listed by ChatGPT itself upfront on entry into its system, some are buried by OpenAI in its fine print, some lie in the OpenAI blog and some limitations are not mentioned at all, but on

examination become self-evident. Yet despite these limitations and caveats both explicitly expressed and implicit in the product, these are largely ignored by naïve Users as ChatGPT is used in ways that go well beyond its capabilities and in fact could lead to serious harm when used by everyday people going about their normal lives.

5 User behaviour in the face of explicit goals, limitations and caveats

5.1 Ignoring plain language notices

On entry into the platform in July 2023, ChatGPT developers state up front, in large print and easy to use language, their key goal, program limitations and caveats for use:

- *"**This is a free research preview"** (bolding theirs not mine)*
- *"Our goal is to get external feedback in order to improve our systems and make them safer"*
- *"While we have safeguards in place, the system may occasionally generate incorrect or misleading information and produce offensive or biased content. It is not intended to give advice"*
- *"Conversations may be reviewed by our AI trainers to improve our systems"*
- *"Please don't share any sensitive information in your conversations"*
- *"May occasionally produce harmful instructions or biased content"*
- *"Limited knowledge of world and events after 2021"*

Yet Users are ignoring these clear plain language notices as well as the notices that sometimes appear in the output itself.

Time is not being spent on checking for inaccuracies or misleading information as can be seen in the use of ChatGPT *to save time* (Duffy, 2023; ABC News, 2023; Johnson, 2023). Professionals in trusted positions share sensitive and private information for ChatGPT to organise for them despite being told that their interactions and input are visible to OpenAI trainers (Johnson, 2023; Purtill, 2023). Some people are even using it as a confidante and advisor (Purtill, 2023) despite being told that these products are not intended to give advice.

5.2 Ignoring notices in the fine print

OpenAI expands on these limitations and caveats on its use and adds others in its Terms of Use[1] and associated Sharing Publication Policy[2] which together include:

a. The need to be at least 13 years old

b. To not violate any person's rights

[1] https://openai.com/policies/terms-of-use updated March 14, 2023

[2] https://openai.com/policies/sharing-publication-policy November 14,2022

c. To not send any personal information of children under the age of 13

d. The fact that the outputs may "*in some situations result in incorrect Output that does not accurately reflect real people, places, or facts. You should evaluate the accuracy of any Output as appropriate for your use case, including by using human review of the Output*."

e. The need to provide legally adequate privacy notices and obtain necessary consents for the processing of personal data

f. The need to indicate in plain language the role of AI in published material including the extent to which the content is AI generated and/or its role in formulating or shaping content.

The degree to which Users are ignoring these notices and terms pervades the internet. While the terms clearly state that children must be at least 13 years old, adults are not only letting their children use the application (Purtill, 2023), but websites are advising on the best ways for children to use it (Bradford, 2023; Maxwell, 2023) even stating "ChatGPT is open to everyone, including children" (Maxwell, 2023).

It is breathtaking, the degree to which these notices and terms are ignored by Users who would normally be concerned with ensuring the accuracy of information, confidentiality, being transparent about their processes and the provision of privacy notices. These include:

1. Lawyers who use the system paying no heed to the need to check the system's outputs as shown by the lawyers who used faulty ChatGPT generated legal research in a court filing later claiming that the program had tricked them (ABC News, 2023) and the CEO of one of Australia's largest law firms claiming that AI tools like ChatGPT save so much time that it could spell the end of the billable hour (Pelly, 2023);

2. Teachers inputting their notes about students and parent-teacher interviews into the system for the LLM to write their final reports for them (Purtill, 2023) in what would appear to be done without written consent, breaching confidentiality for all students and parents, including students under 13;

3. Medical practitioners using the application to summarize patient care and write letters (Johnson, 2023) which has led West Australia's South Metropolitan Health Service (SMHS), which spans five hospitals, to ban use of the AI bot citing concerns with confidentiality. It has also led to calls from the Australian Medical Association for national regulation to control AI in the health sector (Moodie, 2023).

It would appear that the impressive time saving abilities of the technology over-rides those responsibilities that such professionals would normally undertake in everyday situations, such as undertaking careful research in the discharge of their duties, and not telling unauthorised people about a patient's condition or a student's behaviour and academic abilities.

Rather unsurprisingly, I am yet to see anyone on the internet indicate the role that ChatGPT played in formulating or shaping their content.

5.2.1 Significant real-world risks revealed in updates of fine print

In March 2023, OpenAI's alarm at how their LLM (ChatGPT) and other AI software was inappropriately and incorrectly being used was revealed in its updated usage policies (Open AI, 2023) and no doubt, in part, led to their calls for regulation world-wide. Over and above their disallowed usage of their products for illegal, harmful and bad actor purposes, they have highlighted how ChatGPT and other OpenAI products are not able to provide the expert and reliable information for which people have been using the applications by banning the following applications including:

1. Management or operation of critical infrastructure in energy, transportation and water
2. Violation of privacy including unlawful collection or disclosure of personal identifiable information or educational, financial, or other protected records
3. Provision or use of legal or financial advice without a qualified person reviewing the information
4. Diagnosing or providing treatment for health conditions
5. High-risk government decision making for law enforcement, criminal justice, migration and asylum

They further state their requirements for certain uses of their models:

> *"Consumer-facing uses of our models in medical, financial, and legal industries; in news generation or news summarization; and where else warranted, must provide a disclaimer to Users informing them that AI is being used and of its potential limitations.*
>
> *Automated systems (including conversational AI and chatbots) must disclose to Users that they are interacting with an AI system. With the exception of chatbots that depict historical public figures, products that simulate another person must either have that person's explicit consent or be clearly labeled as "simulated" or "parody.""*

It is quite clear that OpenAI does not want to take any responsibility for any of their worldwide publicly released LLM's unreliable outputs and is putting the responsibility for correctly and appropriately using the technology firmly back on the User stating in their Terms of Use that both the User's input and the output content belongs to the User, that the User is responsible for that content and can use it for any purpose as long as the User complies with the Terms of Use.

6 Users and Implicit Limitations

For the OpenAI developers, the goal of this world-wide release of ChatGPT is but one part of a step in their research program for iterative deployment in which improvements are made to *the language model* in stages (OpenAI, 2022). While there clearly are a significant number of Users who use the system in a way which aligns with the developers' goals, equally a significant number of Users of the system have other real-world goals which the system cannot meet. The most of obvious of these is the use of ChatGPT as an efficient time saving *search* tool. However, as the lawyers who ended up filing fake past cases to the court found out, "*it* (ChatGPT) *was not a search engine but a generative language processing tool*" (Bohannon, 2023).

6.1 Lack of understanding about LLMs and their limitations

What appears not to be understood by many Users is how this program works, and that it does not search for and does not understand the content, but if many sample texts use certain words together in certain ways, then it can predict and in essence make-up what follows next. The more that the same information is repeated in the data set, the more reliable the word predictions As (Zhao, 29 June 2023) p1 state, *"In general, LM aims to model the generative likelihood of word sequences, so as to predict the probabilities of future (or missing) tokens"*; tokens being the word, word part, symbol or punctuation mark that comes next.

6.1.1 Language processing in relation to dataset

One of the most serious limitations for public Users of ChatGPT is the suitability and quality of the data in relation to the User's purpose. While ChatGPT states up front that the application has "*Limited knowledge of world and events after 2021*" it is not obvious to the average User that ChatGPT does not access the internet, rather, it has been trained on a specific data set of knowledge existing up to September 2021 with:

- 60% of the data set comprising a filtered version of the Common Crawl of the internet;
- 22% of the data set comprising Webtext2 which comprises the text of all webpages given 3 up-votes or more by Reddit users;
- 8% of the data set comes from Books1 and Books2, two internet-based corpora of books which have not been publicly released and hence it is unclear what types of books they comprise; and
- 3% of the data set comes from Wikipedia (Zhao, 29 June 2023; Brown, 22 July 2020)

While ChatGPT bases its word predictions on an enormous amount of information (still opaque despite the list above), the data set does NOT include a wide range of expertise found in other online books, print books, standards, manuals, specialist data bases, journals available only by subscription and so forth. Clearly, unbeknown to many of the Users, ChatGPT's data set does not include expert knowledge and therefore cannot be used for these purposes.

6.1.2 Fact versus fiction

However, what is not understood by many Users is that even with an expert data set the program focuses on replicating styles of writing rather than any understanding of its data set, inputs and outputs. Hence it can be highly creative and, without any obvious risks to humanity, can generate fiction such as poems, stories and plays. However, ChatGPT is unable to differentiate fact from fiction, is unable to evaluate the quality of the sources on which it bases its word predictions and is unable to determine when further information needed. Further, due to its stylistic capabilities the LLM creates falsehoods, also known as 'hallucinations' such as fake references (Hillier, 2023), falsely naming individuals as having been convicted of a crime, inventing books and studies that don't exist, and providing technical details that don't make sense (Edwards, 2023).

Hence, even if an LLM's responses were based on a more reliable data set, it still could not be relied on for critical real-world decisions as LLMs do not search or browse for content in that data set rather they generate word sequences, as the lawyers referenced above found out and as was flagged by OpenAI when updating their Usage Policies in

March 2023. Therefore, Users should not blindly use any facts suggested by the program for their own written text despite its authoritative style.

One reason for this excessive trust in the output may be in part due to OpenAI's overly simplistic plain language notices and its marketing:

> *"ChatGPT: get instant answers, find creative inspiration, and learn something new. Use ChatGPT for free today. Try on web."*

These faulty claims were in essence acknowledged by Sam Altman CEO of OpenAI who, shortly after ChatGPT's release, tweeted "*ChatGPT is incredibly limited, but good enough at some things to create a misleading impression of greatness. It's a mistake to be relying on it for anything important right now*", later tweeting again "*It does know a lot, but the danger is that it is confident and wrong a significant fraction of the time*" (Edwards, 2023).

Many Users do not appear to understand where the strengths and weaknesses in the program lie. As cited by Edwards (2023), "*"ChatGPT is great for some things, such as unblocking writer's block or coming up with creative ideas," said Dr. Margaret Mitchell, researcher and chief ethics scientist at AI company Hugging Face. "It was not built to be factual and thus will not be factual. It's as simple as that.""*

7 Users' trust and embrace of LLM products

Despite its shortcomings, particularly as a source of factual information as shown above, many schools, professionals from all walks of life, businesses and politicians are embracing ChatGPT for *factual purposes* at an alarmingly rapid rate showing how LLM products can become readily trusted. What becomes abundantly clear from the discussion above, is that when ChatGPT was made available to the public at large, there were and still are many Users interested in testing its bounds and reporting on this. However, equally there were and still are many other naïve Users who instantly trusted the application and incorporated it into their everyday study, work and private lives. This includes the many commentators and influencers who provide incorrect advice on the various ways to use the product, sometimes based on inaccurate information they received from the AI bot itself.

7.1 Misunderstanding technological applied research

Why do so many naive Users readily trust and embrace ChatGPT's outputs? It appears that these Users may have confused technological applied research with market research. They do not appear to understand that OpenAI's research program aimed to improve an LLM; they do not appear to understand the nature of an LLM and how it differs from a search engine; and so, by default, they are unaware that they need to investigate, test and check the system. As far as many Users are concerned, this is simply a product release behind which a multinational company would stand, protect Users' inputs and warrant ChatGPT's outputs. This belief is reinforced by OpenAI's marketers in their Product page which continues supporting and reinforcing many Users' misconceptions about its abilities and how it could be used as shown above.

When academics undertake research which impacts on the humans involved, they are required to get ethics approval, part of which requires obtaining *informed* consent by those participating. Imagine how different the public response and impacts would have been had the researchers actually gained informed consent about this research release: calling for a controlled number of

volunteers, explaining what their research project was about, what an LLM is and how it works, the data set on which the LLM is trained, the participants' role in the project, the limits of the outputs and why they are limited and so on and so forth. It was extremely irresponsible of OpenAI to engage in such research with such real-world impacts without gaining proper informed consent from the people trialling an AI model that is still in development. Having done this world-wide makes it even worse.

In the absence of understanding the nature of the product and the nature of the research project, trust can be *overly* high particularly when the plain language notices and the marketing, as used by OpenAI, underplay the issues.

7.2 Impressiveness of coherently constructed output

With OpenAI's marketing department overstating capabilities, the goal of the release being unclear, and the cautions being ignored, many unsuspecting Users are left finding the LLM highly alluring. The speed with which the LLM can appear to write grammatically correct coherent text just about anything is impressive. Without closely examining the quality of the output, it appears to be able to: provide facts and recommendations; translate languages; write essays, stories, poems and job applications; brainstorm ideas; summarize text; and even provide feedback on your own writing. This impressiveness, coupled with the hype from OpenAI, influencers and commentators, clouds human judgement, increases trust, and creates the impression that AI and LLMs can do anything.

7.3 Human judgement in the face of anthropomorphised outputs

Judgement is further clouded by the 'human face' given to the computer outputs. ChatGPT outputs, through their conversational and impressive language features, sound and feel human, mimicking particular human behaviours, characteristics and emotions in their interactions with Users such as empathy, care and looking after your best interests. While *adults* who log into an LLM know at an intellectual level that the LLMs are not human, they *feel* that they *are* human. Further, due to LLMs' capabilities and tone of output, naïve Users can also feel that LLMs are knowledgeable, intelligent, trustworthy, reliable and authoritative, in essence omniscient in their abilities. For children, this is particularly risky.

These feelings, and ultimately beliefs, are encapsulated in the following quote from Chang (2023), an adult YouTuber with over a million subscribers who, in his ChatGPT training video, advises his audience that they *"just need to think of ChatGPT as a very, very, very smart person in every single topic there is out there. Think of it as like an assistant, as a mentor, as a friend, as a storyteller, as a researcher …you'll be able to make your life so much easier. ChatGPT can help you in all your personal life stuff as well as your business life".*

These feelings go to the core of how naïve Users judge and are influenced by the program's outputs even though the outputs are simply a prediction of word patterns. While trust levels increase with these feeling and judgements, the words in the output are produced from the system with no care, understanding or empathy.

8 Risks arising now and into the future

This preliminary examination of ChatGPT, and some of the issues arising from the way it is used, raises current as well as future risks, not only with the adoption of LLMs like ChatGPT, but also the incorporation of other forms of generative AI in personal lives, business and government.

8.1 Risks from generative AI being released publicly before it is ready

Releasing LLMs and other forms of generative AI while the technology is still in the research stage or before it is ready is highly risky irresponsible behaviour which can have significant impact on society as a whole. This was seen in part by the uncontrolled release of ChatGPT. While the developers view this research as being another stage in the development of a generative AI or LLM system, by default, it also becomes a large uncontrolled social experiment leading to alarm at what is going on, with OpenAI on the back foot updating its terms and conditions and adding notices as unintended uses and consequences happen around them. The public at large, including businesses, government and education, is also left scrambling to address these risks and so needs protection.

It is essential that generative AI is not released widely in the research phase and that the products are subject to thorough regulatory controls with prescribed conditions and standards prior to any release and that any release with human interaction outside of the AI company is subject to the same ethical standards and requirements of human research including ethics approval on the process to be taken, informed consent and appropriate training before the Users themselves train the generative AI.

When generative AI is released publicly before the research is completed and before Users have been thoroughly educated, uptake can be rapid. This can be clearly seen with ChatGPT. Within 2 months of its 'research preview' release, it was estimated to have reached 100 million monthly users "*making it the fastest-growing consumer application in history, according to a UBS study*" (Hu, 2023). According to Hu (2023) this viral launch of ChatGPT will give OpenAI, backed by Microsoft, an advantage in the market.

This is already leading to generative AI products being promoted and sold widely before their capabilities and impacts are understood. When generative AI is released, whether as part of a research project or not, without controls right at the start, it can be highly compelling and can run away from us all, leading to governments, education facilities and businesses simply throwing up their hands and giving in, as there seems no way to stop this train.

This can be seen most starkly in Education. On the 30th of January 2023, ChatGPT was banned in WA public schools from the start of the school year in line with bans other states (Davis, 2023) with South Australia being the only state to not ban the tool at all. However, with private schools using AI there was a fear that private school students would gain advantage (Kovanovic & Dawson, 2023). By May 2023, the WA government lifted the ban, and by July 2023, the federal government released its draft framework for how ChatGPT will roll out in schools due to fear that students will use it anyway (Belot, 2023). As this train runs away, schools and Universities are on the backfoot, needing to rapidly reconsider assessments and how to use the bot constructively, with some Universities even considering going back to paper and pencil exams (Belot, 2023).

Given that commercial imperatives can over-ride concerns about risks, ethical decision making cannot be left to the AI companies and their developers. We don't allow cars on the road before they meet certain conditions of safety and we don't allow medicines to be released before they are tested and deemed safe. Similarly, we shouldn't allow LLMs and other forms of generative AI to be released without meeting certain conditions and prescribed standards.

8.2 Risks to knowledge base

There are two major risks that I believe need addressing in relation to LLMs and AI in general. These relate to LLM and AI generated misinformation, and control over knowledge and information.

8.2.1 Creating misinformation based on limited or faulty outputs

Despite often impressive outputs, misinformation generated by an LLM which is left unchecked by the User, can create further misinformation, risking our knowledge base and ability to make sound decisions even further. This can already be seen with the use of ChatGPT where Users are using the application as a reliable source of information much like an encyclopaedia; even as a reliable primary source of information about the application itself.

For instance, Bradford in March of 2023 asked ChatGPT if ChatGPT is safe for kids to use, clearly believing that this was the primary source for asking about the product itself, just as one would ask Microsoft about one of its products. In other words, she believed she was going to 'the horse's mouth'. However, the output, using its usual method of predicting words, did not state the key points that you would expect to see if you searched the website of the company that produced this product. It did not state this release was part of a research project, that confidential information should not be inputted as their AI trainers access the inputs, and that under its own Terms of Use, children under the age of 13 are not permitted to use the program. Building on an output of half-truths and platitudes, and despite the plain language notice that ChatGPT is "*not intended to given advice*", Bradford (2023), Safewise's Safety and Security Expert, then continued to espouse the wonder and suitability of the tool with kids including that it is free, does not ask for one's age on sign-up, and can be used when a child is lonely to give them someone (or something) to talk to.

This incomplete and, in parts, faulty information provided by ChatGPT, believed by the commentator to be from an accurate primary source, has already entered the internet via her advice as a Safety and Security Expert who many readers would trust. This blog will now become part of the next Common Crawl of the internet and will be picked up by LLMs and so misinformation will grow further resulting in what has been termed in the literature 'the curse of recursion' where the information on the internet increasingly reflects outputs from the LLMs which when scraped up further increases the prediction of faulty information thus diluting the human content (Shumailov et al, 2023 May 31). <u>This creates a serious threat to our knowledge base</u>.

But it is not just faulty LLM information uploaded to the internet that will create problems. Many in the workplace would use or be expected to use such 'time saving' LLMs. This could well result in untrustworthy business or governmental documentation including internal reports with recommendations to decision makers, and untrustworthy external reports and documentation from businesses and governments. This is made worse when recipients of these documents do not know that an LLM or generative AI application to create the document and the extent to which it was used.

<u>It is therefore essential that everyone is educated on each and every AI and LLM product prior to its use to understand the limitations and why they exist, and to empower individuals to question outputs.</u>

It is also essential that it becomes general practice that each and every document include a statement declaring whether or not generative AI or an LLM was used in its production, the extent to which it was used, the reliability of the content and the degree to which the User can or cannot rely on the information.

8.2.2 Control over access to knowledge

The second potential risk to knowledge that I would like to raise is the question of who will control access to reliable sources of knowledge. It is essential that we ensure complete and free access to sources of knowledge. Multinational internet companies are well aware of the value of knowledge and information, and they are already capitalising on this in their search engines, controlling who and what rises to the top of search, and determining who receives what information in their various news feeds and the like.

Those who hold knowledge and control access to knowledge are and will be the winners. This has not been lost on the large players. In 2009, Skidelsky reported on Google's attempt to digitize the contents of all the world's major libraries. At the time of his reporting, Google had already scanned 10 million books leading many to question the motivation behind the scanning. By 2011, Google had digitised about 15 million books and Helft (2011) reported on the company's plan to digitise every book ever published raising concerns "*about the company's growing power over information*".

Ben Lewis's 2013 documentary "Google and the World Brain", is highly revealing about the extent to which multinational internet companies want knowledge control. In it, interviews with Head Librarians reveal how Google staff approached each of them individually and got their permission to digitise the books in their library on the basis of making knowledge, otherwise found only in their library, widely available. Interestingly, at the same time, Google required them to sign confidentiality agreements about this action.

As noted by Newton (2013), the documentary also revealed that "*since 2002 several other corporations have started book-scanning projects of their own, with Microsoft, Amazon, and Baidu among them. And beyond the world of books, companies in every industry are working to amass giant bodies of data, containing personal information about millions of people, over which those same people have little to no control*". Over and above issues of copyright, the documentary further raised the question of whether large companies would restrict access to knowledge or, once in their control, charge for access. We must ensure that this never happens.

We must proactively protect not only our information, but also our sources and our knowledge base, and ensure free access to that knowledge and information.

8.3 Risks from incorrect use of AI

Another significant risk that needs addressing is people unwittingly using an AI program for purposes for which it is simply not capable of doing and was never designed to be able to do, irrespective of whether it is still in the research phase or not. We can see this in the incorrect uses to which ChatGPT is being put by a significant number of Users. These misuses of ChatGPT by everyday citizens caught even OpenAI by surprise causing them to have to make rapid changes to address some of the issues.

Some of these changes enacted by OpenAI could be readily seen by Users, such as incorporating an ability to turn off the chat history so that confidential information cannot be used by AI Trainers

in April 2023 (OpenAI, 2023). Other changes, however, remained hidden, such as OpenAI changing their Usage Policies in March 2023 (Open AI, 2023) to disallow uses for which ChatGPT is simply incapable of doing as discussed earlier. While the Usage Policy changes may aim to protect OpenAI legally, remaining buried in the fine print does nothing to protect society at large. Unwittingly, many will continue to engage in such banned uses which could result in significant risks to human safety, health, financial well-being, and justice, to name but a few.

It is not acceptable for a company to simply update their Terms and Conditions. We must provide clear parameters for the actions AI companies need to take, including thorough educational programmes (i) to mitigate against misuse and (ii) to address misuse when it occurs.

8.4 Risks from blind acceptance of AI and LLM outputs

As outputs from LLMs and AI improve over time, and as the opacity which underlies the output increases, so the belief in AI's omniscience will most likely increase within the population at large. Hence people will feel that they are not able to disagree with or challenge an LLM or AI output no matter what position they are in: client, recipient of a service, service provider, boss, employee – to name but a few.

We have already seen this type of behaviour with Robodebt, where Automated Decision Making (ADM) was used to send out debt notices to social security recipients. Recipients of the Robodebt letters felt like they were hitting a brick wall when they questioned the decision (SBS News, 2019). According to Michelle Lazarus and Joel Townsend of Monash University (2023) so many people trusted the system because the decision making was seen to be objective and trustworthy. Further, it removed human uncertainty causing automation bias by departmental personnel as well as some of the recipients of the notices.

We must make sure that it does not happen again by ensuring that all Users are educated about the technology they are using, and by ensuring practices that will retain human control along with the skills, knowledge and ability to challenge.

8.5 Risk of increased distance between businesses and clients

It is of grave concern that as LLMs become increasingly used by businesses and governments as the interface between the organisation and the public, so the distance between organisations and their clients will increase. We can already see this distance in the use of call centres where the call centre is the only point of contact with the organisation, the call centre staff are often not the decision makers and not empowered to make decisions, and where it is almost impossible to get through to anyone who does make the decisions. Often, we cannot even find the names let alone the contact details of any person within a department. It never used to be so. This distance coupled with blind acceptance of AI and LLM outputs will further exacerbate problems. Actions need to be taken to ensure that the public can always speak to a human who is skilled in the matters to do with the organisation, empowered to make decisions and to investigate situations arising.

8.6 Risk of deskilling of the population

Finally, while we may aim to for human oversight over all AI and LLM enabled activities, there is a significant risk that leaving the thinking and decision making to machines will result in significant deskilling of the population with generations of humans progressively being unable to evaluate or

judge the quality of AI and LLM outputs, let alone be able to feel that they can be more insightful, accurate and reliable than a machine.

It is essential that we continue to develop and retain skills and knowledge that we currently have in order to be able to provide that important human oversight. This includes the ability and skills to search for information, organise information, analyse information, make judgements and predictions within a given discipline area, field of research, profession or job of any type without needing to rely on AI.

9 References

ABC News. (2023, June 9). *Lawyers in the United States blame ChatGPT for tricking them into citing fake court cases*. Retrieved from ABC News: https://www.abc.net.au/news/2023-06-09/lawyers-blame-chatgpt-for-tricking-them-into-citing-fake-cases/102462028

Bellan, R. (2023, March 14). *Microsoft lays off an ethical AI team as it doubles down on OpenAI*. Retrieved from TechCrunch: https://techcrunch.com/2023/03/13/microsoft-lays-off-an-ethical-ai-team-as-it-doubles-down-on-openai/

Belot, H. (2023, July 9). *ChatGPT ban in Australia's public schools likely to be overturned*. Retrieved from The Guardian Austrlaian Edition: https://www.theguardian.com/technology/2023/jul/09/chatgpt-ban-in-australias-public-schools-likely-to-be-overturned

Bohannon, M. (2023, June 8). *Lawyer Used ChatGPT In Court—And Cited Fake Cases. A Judge Is Considering Sanctions*. Retrieved from Forbes: https://www.forbes.com/sites/mollybohannon/2023/06/08/lawyer-used-chatgpt-in-court-and-cited-fake-cases-a-judge-is-considering-sanctions/?sh=57a517767c7f

Bradford, A. (2023, March 17). *What Is ChatGPT and Is It Safe for Kids?* Retrieved from Safewise: https://www.safewise.com/what-is-chatgpt-and-is-it-safe-for-kids/

Brown, T. et al (22 July 2020). Language Models are Few-Shot Learners. *arXiv:2005.14165v4*.

Chang, C. (2023, April 4). *ChatGPT Tutorial: How to Use Chat GPT For Beginners 2023.* Retrieved from Youtube: https://youtu.be/Gaf_jCnA6mc

Davis, A. (2023, January 30). *ChatGPT banned in WA public schools in time for start of school year*. Retrieved from ABC News: https://www.abc.net.au/news/2023-01-30/chatgpt-to-be-banned-from-wa-public-schools-amid-cheating-fears/101905616

Duffy, C. (2023, May 26). *Public school bans on AI tools like ChatGPT raise fears private school kids are gaining an unfair edge and widening a digital divide*. Retrieved from ABC News: https://www.abc.net.au/news/2023-05-26/artificial-intelligence-chatgpt-classrooms-schools/102356926

Edwards, B. (2023, June 4). *Why ChatGPT and Bing Chat are so good at making things up*. Retrieved from Arstechnica: https://arstechnica.com/information-technology/2023/04/why-ai-chatbots-are-the-ultimate-bs-machines-and-how-people-hope-to-fix-them/

Helft, M. (2011, March 22). *Judge Rejects Google's Deal to Digitize Books*. Retrieved from The New York Times: https://www.nytimes.com/2011/03/23/technology/23google.html

Hillier, M. (2023, Feb 20). *Why does ChatGPT generate fake references?* Retrieved from Teche Macquarie University: https://teche.mq.edu.au/2023/02/why-does-chatgpt-generate-fake-references/

Hu, K. (2023, February 2). *ChatGPT sets record for fastest-growing user base - analyst note*. Retrieved from Reuters: https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/

Johnson, K. (2023, April 24). *ChatGPT Can Help Doctors—and Hurt Patients*. Retrieved from Wired: https://www.wired.com/story/chatgpt-can-help-doctors-and-hurt-patients/

Kovanovic, V., & Dawson, S. (2023, July 6). *High school students are using a ChatGPT-style app in an Australia-first trial*. Retrieved from The Conversation: https://theconversation.com/high-school-students-are-using-a-chatgpt-style-app-in-an-australia-first-trial-209215

Lazarus, M., & Townsend, J. (2023, March 22). *Automation, uncertainty, and the Robodebt scheme*. Retrieved from Monash University Lens: https://lens.monash.edu/@michelle-lazarus/2023/03/22/1385582/automation-uncertainty-and-the-robodebt-scheme

Loizos, C. (2023, March 29). *1,100+ notable signatories just signed an open letter asking 'all AI labs to immediately pause for at least 6 months'*. Retrieved from TechCrunch: https://techcrunch.com/2023/03/28/1100-notable-signatories-just-signed-an-open-letter-asking-all-ai-labs-to-immediately-pause-for-at-least-6-months/

Maxwell, T. (2023, March 11). *5 Ways Kids Can Use ChatGPT Safely*. Retrieved from Make Use Of: https://www.makeuseof.com/ways-kids-can-use-chatgpt-safely/

Moodie, C. (2023, May 28). *Australian Medical Association calls for national regulations around AI in health care*. Retrieved from ABC News: https:://www.abc.net.au/news/2023-05-28/ama-calls-for-national-regulations-for-ai-in-health/102381314

Newton, C. (2013, Jan 25). *Documentary throws the book at Google scanning project* . Retrieved from CNET : https://www.cnet.com/tech/services-and-software/documentary-throws-the-book-at-google-scanning-project/

Open AI. (2023, March 23). *Usage Policies*. Retrieved from Open AI: https://openai.com/policies/usage-policies

OpenAI. (2022, November 30). *Introducing ChatGPT*. Retrieved from OpenAI: https://openai.com/blog/chatgpt

OpenAI. (2023, April 25). *New ways to manage your data in ChatGPT*. Retrieved from OpenAI: https://openai.com/blog/new-ways-to-manage-your-data-in-chatgpt

Pelly, M. (2023, April 27). *Minters boss tips ChatGPT to end billable hour*. Retrieved from The Australian Financial Review: https://www.afr.com/companies/professional-services/minters-boss-tips-chatgpt-to-end-billable-hour-20230419-p5d1o8

Purtill, J. (2023, April 15). *How Australians are using ChatGPT and other generative AI in their everyday lives*. Retrieved from https://www.abc.net.au/news/science/2023-04-15/australians-using-generative-ai-everyday-life/102214676

SBS News. (2019, Oct 9). *'I don't have any trust': Centrelink robo-debt recipients say it feels like bullying*. Retrieved from SBS News: https://www.sbs.com.au/news/article/i-dont-have-any-trust-centrelink-robo-debt-recipients-say-it-feels-like-bullying/4lrs5bkw8

Shumailov, I. et al (31 May 2023). The curse of recursion: Training on generated data makes models forget. *arXiv:17493v2*.

Skidelsky, W. (2009, Aug 30). *Google's plan for world's biggest online library: philanthropy or act of piracy*. Retrieved from The Observer: https://www.theguardian.com/technology/2009/aug/30/google-library-project-books-settlement

Zhao, W. X. et al (2023, June 29). A Survey of Large Language Models. *arXiv:2303.18223v11*.