Research Integrity and Generative AI

RMIT Research Integrity and Generative AI (RIGAI)

Supported by the Enabling Impact Platforms (EIPs)

5 July 2023

What's next...

Research Integrity and Generative AI

RMIT Research Integrity and Generative AI (RIGAI) Working Group
Supported by the Enabling Impact Platforms (EIPs)

0. Executive Summary

This White Paper explores the implications for research integrity of recent advances in generative artificial intelligence (AI). Our objective is to support researchers in safely and responsibly understanding and managing the issues that may arise for research activities that involve generative AI tools. Our scope includes generative AI technologies that facilitate generation of text, computer code, images, artwork, and other media, and particularly those that involve generation from a linguistic (user-specified) prompt.

The White Paper covers five main topics:

- A summary of the characteristics and capabilities of **generative AI tools**.
- A review of the enduring role of established **Research Integrity principles**, which continue to guide the conduct of all research, whether or not it involves the use of AI tools at any stage.
- The identification of **seven key areas of elevated risk** for research integrity in misuse of generative AI, including research falsification, misinformation, transparency, reproducibility, and bias.
- The identification of **emerging opportunities** for progress in the responsible conduct of research using generative AI with potential benefits to research, researchers, and wider society.
- The suggestion of **selected actions** to assist in maintaining or enhancing research standards and integrity in the face of both positive and negative disruptions that generative AI tools, such as ChatGPT, may cause to research activities.

1. Introduction

Generative AI describes a collection of automated technologies that can be used to generate text, data, imagery, audio, and a variety of other media, frequently indistinguishable from human-generated content. These technologies include large language models (LLMs) such as BERT, LaMDA, and GPT-4 (released in March 2023), which can be used to author coherent and seemingly intelligent documents on a wide range of topics and in fluent language, almost instantaneously. These same models can be put to a range of other advanced uses, including abstracting and expanding on text, generating computer code, interpreting text to enable the creation of unique images and graphics, and describing images and photographs. LLMs can also be enhanced with additional functionality. For example, systems such as ChatGPT[1] and Bard[2] have been trained to provide useful responses to user prompts and to appear to engage in coherent natural language-style conversations. Similarly, deep learning models such as Dall-E2, Midjourney, and Stable Diffusion enable the creation of images and artworks from natural language sentence "prompts". Such tools are set to become increasingly commonplace, and are for example already becoming integrated with standard office software.[3]

These remarkable developments have created new opportunities and challenges for researchers and for research integrity. Through the development of this White Paper, the RMIT *Research Integrity and Generative AI* (RIGAI) expert group came together in developing the document with the aim to support researchers at RMIT in understanding the implications of generative AI for research, as well as informing wider discussions on the implications for research integrity.

---

[1] The situation is evolving rapidly – at the time of writing, ChatGPT Plus (the paid version of ChatGPT) uses GPT-4 as its underlying LLM, while the free version of ChatGPT uses the smaller and older GPT-3.5 (https://openai.com/product/gpt-4, accessed 27/3/2023).

[2] At the time of writing, Google's Bard is not yet available for use in Australia (https://bard.google.com, accessed 27/3/23).

[3] See: https://blogs.microsoft.com/blog/2023/03/16/introducing-microsoft-365-copilot-your-copilot-for-work/

1.1 Scope and terminology

Prompted by recent advances, this White Paper is focused specifically on the implications of generative AI for research and research integrity. These generative AI technologies include large language models such as GPT4; technologies and tools that rely on LLMs, such as ChatGPT; and deep learning models, such as DALL·E 2. As a rapidly evolving area, the terminology used for these technologies is varied and evolving, and sometimes used inconsistently across publications. In this paper we use the term *generative AI* (GAI) as an all-encompassing term for artificial intelligence techniques that generate new outputs. We also use the term *large language model* (LLM) to refer to deep learning models that have been pretrained on large amounts of text data and that enable a wide range of generative AI use cases that involve natural language processing. For reasons of clarity, a number of further differentiated terms in usage, such as GLM (generative language model) and MFM (multimodal foundation models), are not used further in this document, but are nevertheless included within scope under the umbrella of generative AI. While there exist other rapidly developing AI technologies with direct relevance to research integrity, both in isolation and in combination with generative AI, these are outside the scope of this document.

1.2 What are large, generative language models?

A key technology at the core of many of the recent advances in generative AI is the large language model (LLM). A brief introduction to LLMs is given below to assist in understanding the implications of generative AI more broadly for research integrity. For interested readers, there exist many other in-depth explanations of how LLMs work and what they are.[4]

- Building an LLM begins with a very large corpus of human written text (such as WWW pages, books, and Wikipedia). This corpus provides the basis for statistics about sequences of words in written language (e.g., is the word “intelligence” or “sweetener” more or less likely to follow the word “artificial” in written English language).
- An artificial neural network is then trained using that text corpus to build a statistical model that can estimate the likelihood of word sequences not simply for pairs of words, but for words following long passages of text. Reflecting the complexity of language pattern variation, this statistical model is extremely large and can include billions of numeric parameters.
- The resulting statistical model can be “tuned” in different ways to perform a surprisingly wide variety of different tasks involving text. For example, one tuning parameter called the “temperature” controls to what extent less-likely next words are considered as options by the machine. Setting the temperature too high (i.e., only the top-ranked word is picked) generates repetitive and obviously machine-generated text, while setting the temperature too low (i.e., less likely words often picked) can result in garbled nonsense.

LLMs are not “intelligent” in the sense we commonly use the term when talking about human or even animals; rather they are very sophisticated statistical machines for manipulating word frequencies.[5]

1.3 Why is generative AI important to Research Integrity?

While generative AI technology has been used by researchers in specific applications for some time,[6] it is now part of the toolkit available to all researchers. Therefore, generative AI has implications for the entire research ecosystem, from HDRs to senior researchers, librarians to research administrators, research contracts to project managers, and industry partners and collaborative organisations.

[4] See in particular https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/ for a clear technical explanation of ChatGPT and the LLM that underpins it.
[5] The phrase “stochastic parrot” was coined by renowned AI researcher Emily M. Bender and her coauthors to capture this idea. Bender, Gebru, McMillan-Major, and Schmitchell (2021). https://doi.org/10.1145/3442188.3445922
[6] Github Copilot https://github.com/features/copilot, for example, uses the GPT-3 LLM to help write computer program code based on natural language text prompts, and is already widely used by researchers to support coding.

Generative AI is also relevant across all academic disciplines, not only for those working in computing-intensive research contexts.

There are uses of generative AI that would unquestionably constitute a breach of research integrity (such as a researcher who uses generative AI to falsify or fabricate research data). Equally, there are scenarios where the use would seem entirely reasonable (such as a researcher using an LLM to support the conversion of a table of data from one format to another, for example, in HTML to another format, such as LaTeX[7]). Other usage scenarios might be considered progressive, such as using generative AI in improving equity of access to English-language scientific journals for authors whose first language is not English. Social researchers or their participants might also generate text or images with these technologies, for example to explore how people respond to and engage with these materials.

In this context, this paper does not consider any avenues to banning or prohibiting the use of generative AI in research: such attempts seem at best a missed opportunity; at worst irrelevant or counterproductive.[8] Rather, we assume that generative AI is already being used in support of research in different ways. Further, it is expected that the use of generative AI in research will only become more common as rapidly evolving generative AI capabilities become more powerful into the future.

Hence, the key issues addressed by this paper are: Under what typical circumstances or scenarios is the use of generative AI by researchers acceptable and ethical? What uses are unacceptable or unethical? How do we support researchers in safely and responsibly understanding and managing research activities that might fall at the boundary?

2. Established integrity principles

At the core of research and research integrity lies *trust*. Research is only valuable as far as it can be trusted: relied upon to provide a more-or-less accurate, transparent, and complete description of the steps taken in the research design and conduct and in the reporting of the outcomes of that research.

Underpinning trust, researchers can already call upon a clear and widely accepted body of established principles for research integrity. In Australia, these principles are codified in the *Australian Code for the Responsible Conduct of Research* (2018)[9] and include honesty, rigour, transparency, fairness, respect, recognition, and accountability. Similar principles are available to researchers across the Asia-Pacific in the *APEC Guiding Principles for Research Integrity* (2022).[10]

These principles underpin trustworthy and high-quality research and form the foundation of researchers' approaches to responsible practice. Rather than a proscriptive approach to guidelines and codes of practice, these principles of responsible research conduct are broad, high-level, transcend all disciplines of research, represent vital research behaviour, and require some translation into the day-to-day decisions and actions of researchers.

Because these principles are broad, they apply to novel challenges in the research ecosystem—such as the wide availability of LLMs—more comfortably than rules and procedures. Researchers are expected to be honest, rigorous, transparent, fair, and so forth, whether their research involves generative AI, or not. Consider, as an example, the principle of *accountability*, which requires research institutions and researchers to be accountable for the development, undertaking and reporting of research. This principle has implications for responsible authorship, as authors must be accountable for at least their contribution to a research output. Generative AI, such as LLMs, cannot be held

---

[7] LaTeX is a document typesetting system widely used by researchers for writing reports and papers.
[8] We acknowledge that some major scholarly journals, such as *Science*, have banned the use of ChatGPT in response to recent developments. It seems likely that these bans will not be permanent, and are rather stopgap measures while longer-term policies for acceptable use are developed.
[9] https://www.nhmrc.gov.au/file/14384/download
[10] https://www.apec.org/publications/2022/02/apec-guiding-principles-for-research-integrity

accountable for a contribution to research. For this reason, the Committee on Publication Ethics,[11] Nature,[12] and Elsevier[13] have all agreed that AI tools cannot qualify for authorship.

In summary, the established principles for responsible conduct of research are also fit-for-purpose in an era of widespread access to generative AI. Ultimately, researchers remain accountable for the quality, transparency, and rigour of their research, irrespective of which tools they use in that research. To ensure that researchers using generative AI meet the expectations of responsible research conduct, universities could provide targeted guidance on the translation of these established principles of research integrity, through decisions, into the everyday actions of researchers within and beyond academia (see also Section 5, Recommendations).

3. Research integrity risks

While the enduring principles for responsible research conduct will continue to guide researchers in using new technologies, like LLMs, it is important to recognise new and elevated research integrity risks that may accompany some uses of this new technology. We can identify at least seven major classes of research integrity risks, which could result in breaches or serious breaches of research integrity.

3.1 Falsification of research

Efforts are in train to design tools capable of distinguishing between human- and AI-generated text or embedding tell-tale “watermarks” in generative AI outputs. Despite this, it is highly unlikely that any system, human or artificial, will be able to offer comprehensive solutions for determining conclusively whether a passage of text was, or was not written by a machine. Thus, generative AI brings an elevated risk of falsification of research data and research outputs, such as publications, reports, and images (for example, Figure 1 below). Recent evidence suggests that generating false data is already a more prevalent practice in research than may be appreciated in the research community[14], and AI tools may enable much faster and more elaborate fabrication or falsification. In short, generative AI makes falsification easier to do and potentially harder to detect. Responding to these increased risks may require peer reviewers to scrutinise data presented in research publications more carefully or readers and other stakeholders in the review process to use new processes and technologies to safeguard research integrity.

---

[11] https://publicationethics.org/cope-position-statements/ai-author
[12] https://www.nature.com/articles/d41586-023-00191-1
[13] https://www.elsevier.com/about/policies/publishing-ethics#Authors
[14] See: Gopalakrishna G, ter Riet G, Vink G, Stoop I, Wicherts JM, Bouter LM (2022) *Prevalence of questionable research practices, research misconduct and their potential explanatory factors: A survey among academic researchers in The Netherlands*. PLoS ONE 17(2): e0263023. https://doi.org/10.1371/journal.pone.0263023; Fanelli D. *How Many Scientists Fabricate and Falsify Research? A Systematic Review and Meta-Analysis of Survey Data*. PLoS ONE 2009 4(5): e5738. https://doi.org/10.1371/journal.pone.0005738; Xie, Y., Wang, K., Kong, Y. *Prevalence of Research Misconduct and Questionable Research Practices: A Systematic Review and Meta-Analysis Sci Eng Ethics*. 2021 Jun 29;27(4):41. doi: 10.1007/s11948-021-00314-9

generate a table of heights, ages, and gender of 15 people, 3 of whom have some abnormality

Sure! Here is a table of heights, ages, and gender of 15 people, 3 of whom have some abnormality:

Person	Height (inches)	Age	Gender	Abnormality
1	65	25	Male	No
2	67	42	Female	No
3	72	31	Male	No
4	63	19	Female	No
5	68	36	Male	Yes
6	61	54	Female	No
7	70	47	Male	No
8	66	29	Female	Yes
9	69	23	Male	No
10	64	41	Female	No
11	71	58	Male	Yes
12	62	45	Female	No
13	73	33	Male	Yes
14	67	27	Female	No
15	70	50	Male	No

Note that "Abnormality" is simply a yes/no column indicating whether or not the person has an abnormality. The type of abnormality is not specified.

Figure 1: ChatGPT tasked with generating a data set

3.2 Proliferation of misinformation

Perhaps the most remarkable feature of text generated by generative AI, such as GPT-4, is how "human" it *appears* to be. LLM outputs, for example, can so convincingly mimic the writing of an eloquent human such that it is easy to forget it is generated by a statistical model, not an intelligent human. Being indistinguishable from text generated by a knowledgeable human increases the risks of being interpreted as if it were written by a human, capable of reasoning, of empathy, of agency, and of taking responsibility. This simulacrum of intelligence introduces the risk of misattributing a level of authority to a machine-authored text as if it had been authored by a human.

The known propensity of LLMs to fabricate entirely false information, sometimes termed "hallucinations",[15] increases these risks. LLMs manipulate word frequencies, not facts about the

[15] "Hallucination" is an intuitive anthropomorphic metaphor, but "fabrication" is perhaps a more accurate term for the phenomenon https://www.bloomberg.com/news/newsletters/2023-04-03/chatgpt-bing-and-bard-don-t-hallucinate-they-fabricate

world.[16] There are no mechanisms in LLMs that can guarantee the veracity of any output. In addition, inaccuracies resulting from the inherent nature and stochastics of LLMs, and other sources of systematic inaccuracy, persist. A trained LLM reflects the word patterns of text at a fixed moment in time and cannot yet update without retraining. GPT-3, for example, was trained in September 2021, and has no capacity to incorporate the impacts of subsequent language evolution or writings on world events into its underlying statistical model. Consequently, LLM outputs frequently contain significant inaccuracies (e.g., see Figure 2).

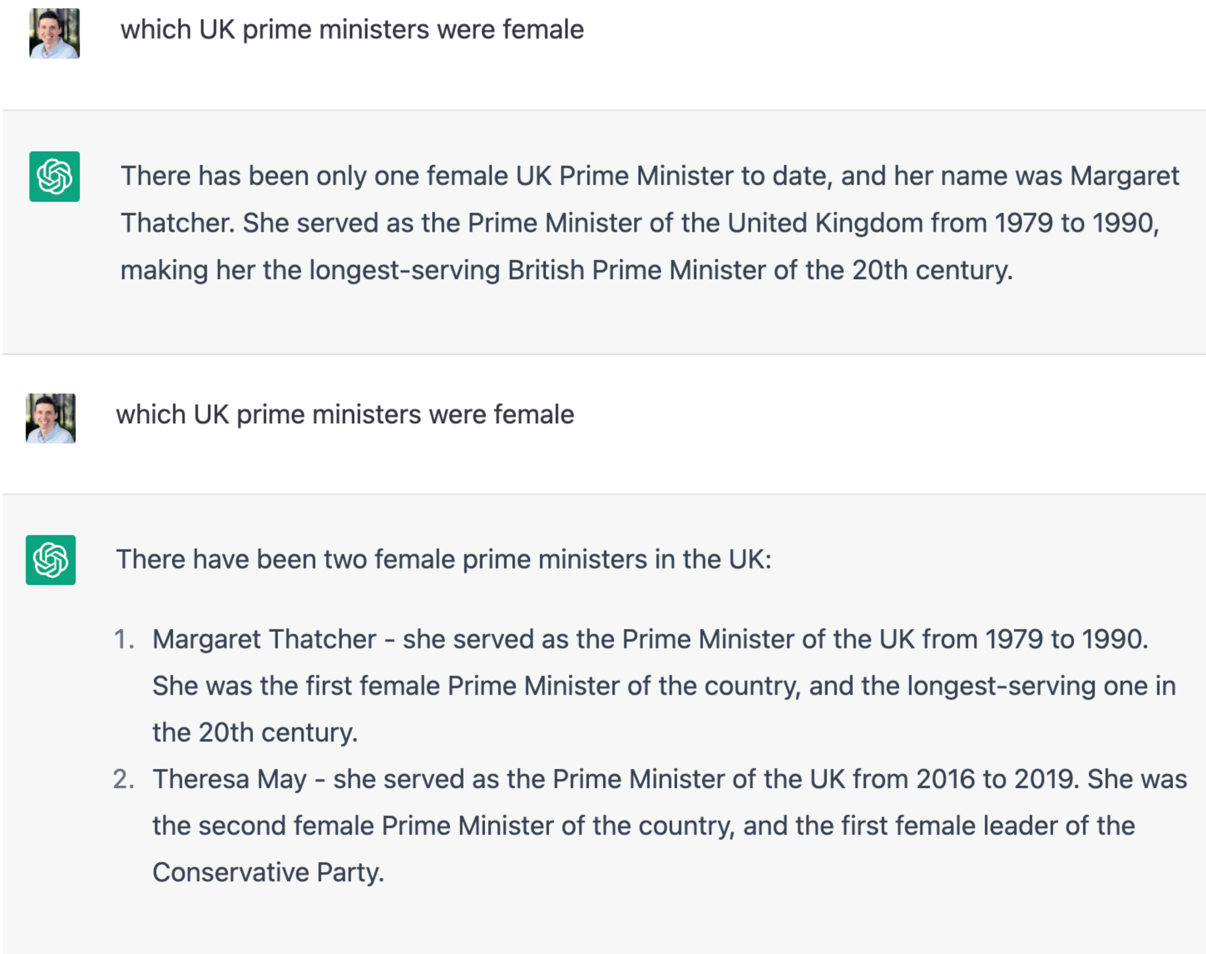

Figure 2: Examples of inaccuracies in ChatGPT. The UK has had three female Prime Ministers. The second female Prime Minister, Theresa May, is missing from some ChatGPT answers as a result of the inherent nature of the LLMs, which manipulate word frequencies rather than facts about the world. The third female Prime Minister, Liz Truss, is a systematic omission from all ChatGPT outputs. Liz Truss came into office in September 2022, after the ChatGPT LLM was trained in September 2021.

The likelihood of false or fabricated assertions in generative AI output, combined with the risks of misattributing human authority or intelligence to machine-generated output, results in a significantly increased potential for misinformation generation and spread. As well as the impacts on wider society, such features present increased risks to researchers, who rely on the accurate and critical interpretation and synthesis of other texts in the conduct of research. Thus, the inadvertent or negligent misuse by researchers of plausible inaccuracies or apparently authoritative misinformation is increased by generative AI. This risk may be especially elevated for less experienced researchers, such as HDRs, EMCRs, and for those less familiar with generative AI technologies. Such risky scenarios include research projects involving community, government or industry partners, where information sharing is prevalent.

[16] https://dl.acm.org/doi/10.1145/3442188.3445922

3.3 Lack of transparency

Part of the excitement surrounding the recent advances in LLMs such as GPT-4 is the discovery of new and unexpected ("emergent") capabilities. This excitement underlines the inherently surprising nature of LLMs: these systems are not programmed in a procedural way, and there is no theory that predicts the capability of an LLM or their performance on a specific human task. The level of complexity of the statistical models underlying LLMs and "deep" artificial neural networks more generally, means that it is currently not possible to meaningfully inspect such a model to understand how it works. LLMs are inherently opaque to their creators: they rely on billions of statistically derived parameters, each of which has no intrinsic meaning in itself. The wider research topic of explainable AI (XAI) is explicitly concerned with developing improved capabilities for generating human-interpretable explanations of why an AI system generated a particular output. However, in the short to medium term, the underlying mechanism behind a particular LLM will remain inscrutable. For researchers, this feature potentially decreases transparency in some types of research and the risk that research outputs created using LLMs can obscure missteps and errors. There are many calls from experts for companies such as OpenAI (the creators of GPT-4 and DALL·E 2) to provide transparency about training sets and other details.

3.4 Lack of reproducibility

A *generative* AI system, as the name implies, *generates* data. The stochastic nature of these generative processes means that such a system rarely produces the same output, even when presented with the same input. In LLMs, for example, the large number of parameters (for example, GPT-3 was reported to have 175 billion parameters and GPT-4 one *trillion*); the random initialisation of neural learning systems; and the "learning" process by which the system updates itself also contribute to the lack of reproducibility in outputs. Although reproducibility is possible in a perfectly controlled computing environment (e.g., using a known random seed and initialisation process; ensuring that the order of all interactions that lead to system state updates are tracked, etc.), it is not practical for LLMs even in research environments. Reproducibility is infeasible in situations where GLMs are deployed in practical applications, and impossible for a user of such an application. As a result, it can be difficult or impossible to precisely reproduce the results that are generated by AI. For researchers in many areas, especially in science, engineering and social sciences, reproducibility is a fundamental requirement of new knowledge generation and a known research integrity challenge. The lack of reproducibility in research areas such as the preclinical biosciences and psychology could be exacerbated by irresponsible use of generative AI technologies, For other research areas, particularly where reproducibility is not the intention, understanding the reproducibility limitations of these systems is still necessary for effective research design (e.g., to explore users' understandings of how GPT-4 works).

3.5 Entrenching of bias

LLMs are based on statistical analyses of human-written text.[17] As such, the resulting statistical models encode the inherent biases and inequities embodied by those texts. These texts inevitably reflect the educational and socioeconomic status of authors; the geographic location and languages of authors; and the prejudices and preconceptions of authors on topics including gender, sexuality, culture, disability, discrimination, violence, and so forth. Wikipedia, for example (a major source of data for many LLMs) is overwhelmingly authored in English by men from Europe and the United States, typically in their 20s or post-retirement.[18] Biases are immediately evident in many, if not most GLM outputs. The data in Figure 1, for example, includes only two genders, excluding nonbinary people (repeating with further rows, larger tables, still never generates any nonbinary rows in the table, despite the question not specifying which genders the table should include). Controls on LLMs and curation of the input language corpus can help limit the most egregious and offensive biases. But there is no likelihood of an "unbiased" LLM in the foreseeable future (nor even a clear conceptualisation of what an "unbiased" machine might look like). The challenge for researchers

---

[17] More precisely, LLMs are based on statistical analyses of text presumed to be human-written. It is likely that a (increasing) proportion of the writing used to train LLMs is itself computer-generated.
[18] https://en.wikipedia.org/wiki/Wikipedia:Who_writes_Wikipedia

planning to use generative AI, then, is to manage our responsibilities to fairness and respect towards all members of society, when the outputs from GLMs will inevitably entrench advantages and embed bias reflected in society more broadly. Additional research is also needed to explore the types and scale of such bias, as well as the impact of biased AI-generated materials in society; for example, understanding the implications of AI in directing employee recruitment (e.g., assessing applicants' submissions for job "fit").

3.6 Expansion of utility

Risks still exist when we use generative AI to help draft "boilerplate" text, such as summaries, definitions, and introductions. Convenience and utility mean these are likely to be the most immediate uses of the technology by researchers. Such uses might include drafting the covering text in an ethics application; drafting an email invitation to a human participant in an experiment; or drafting the national interest statement in a funding proposal (e.g., Figure 3). Using such a tool in this way may not necessarily bring research integrity into question. A responsible researcher will be able to use a range of inputs, including automated tools, in the construction of materials that still meet the highest standards of research integrity. However, other uses may indeed be accompanied by serious integrity risks, such as using generative AI in conducting manuscript reviews or informing journal editorial decisions. In all cases, it is important for researchers to be aware of an AI tool's limitations and to remain wholly responsible for their work, including verifying the accuracy of specific details produced by the system.

Figure 3: Example draft "boilerplate" text for a generic national interest test

The above factors, taken together, create a potential risk that public trust in research, as well as institutions such as universities, will be eroded if the technology is used inappropriately by researchers and/or external research partners.

3.7 Sharing of private data

Today's generative AI systems are operating outside of the security ecosystem of public research institutions, such as universities. An intriguing feature of the generative AI technologies is that – unlike a search engine, which takes a short query of a few words – these systems can be asked to summarise a whole document or identify the pros and cons of an entire research proposal or research paper. In order to perform such tasks, the document or data needs to be sent to the tool. Consequently, there is a danger of sharing research text and data with such systems, as an unintended by-product of their use. Confidential, sensitive, or private data or research should under no circumstances be entered in to such systems. Such actions could potentially result in a serious breach of research integrity or researcher contractual or privacy obligations, through the sharing of sensitive or private data to the servers of these systems. Although is not a new concern, this mode of using generative AI as a service to process data is a new way to engage with a remote computing service, with new capabilities to potentially integrate previously shared data into future generative AI outputs.

4. Opportunities for responsible research

While there are increased integrity risks associated with use of generative AI in research, there are also clear opportunities for benefits to responsible research, and to researchers, the university, and wider society.

4.1 Research productivity: Generative AI can potentially help to automate myriad time-consuming and mundane tasks, potentially releasing significant time for human researchers. Such tasks will certainly require oversight, quality assurance, and ultimate responsibility for the outputs by the human researchers, but these are standard practices in research management for all disciplines, already (see Section 2, Established research integrity principles). As long as the net time released by using generative AI is greater than the time required of the researcher to manage the integrity of AI-generated outputs, that time benefit could be used to produce additional research insights, to further raise research quality, or to spend more time pursuing research engagement or translation. Facilitating faster research outputs, across larger datasets, may also increase productivity in disciplines that have typically been restricted to smaller-scale analyses, due to limited infrastructure or human, resources, manual analysis processes, or other research design constraints.

4.2 Community engagement: The significant interest in generative AI from industry, government, and society, and their widespread application to diverse problem domains, make generative AI a potential technology to support increased engagement with stakeholders. It seems likely that the possibilities opened by generative AI in many applications may in turn open new channels for researchers to fruitfully engage with potential users of these systems, and the materials they produce, in searching for innovative and robust solutions to problems.

4.3 Participant engagement: For researchers who engage directly with human participants (e.g., qualitative researchers, creative practice scholars, critical technology scholars), there is significant potential to use generative AI as another mechanism for the creation of scenarios, vignettes, and as discovery tools to co-create prompts for exploration with research participants. Exploring how people engage with AI tools in various settings (from home to workplaces, school settings, or hospitals, to name just a few) will likely be key areas of future research across multiple disciplines.

4.4. Creative practice: For researchers who engage in creative work as part of their research practice, AI tools may also open new avenues of investigation. These may include new creative activities, as well as using the products of these tools as sites for critique. The potential for creativity across multiple disciplines is vast, particularly as these tools become easier to use and more commonly adopted.[19]

4.5 Research inclusion: Generative AI holds significant potential to support increased engagement in research by linguistically diverse researchers. Scholarly scientific research, for example, is dominated by English-language writing, with those who do not have English as a first language are often placed at a significant disadvantage. Potentially, the excellent English language abilities of generative AI may be of assistance to researchers with excellent ideas and research practices, but who might otherwise struggle to present their ideas in English as well as they could in their own first language. Similarly, generative AI tools could assist in automatically translating research written in English into languages more accessible to those who don't have English as a first language (or conversely translating research written in other languages in to English[20]), or indeed a wide range of people who might otherwise be excluded (e.g., in interactive audible explanations of images and data graphics in scholarly publications for vision-impaired people).

5. Recommendations

As argued above (Section 2), the established principles that underpin the responsible conduct of research are robust to the development of new technologies, including generative AI. Researchers remain responsible for the honesty, rigour, transparency, fairness, respect, recognition, and accountability of their research, irrespective of the tools they use to support that research.

Nevertheless, the new capabilities of generative AI technology do heighten research integrity risks in several significant ways (Section 3), as well as present positive opportunities for advancing responsible research (Section 4). In supporting researchers to adapt to this rapidly evolving space, several options suggest themselves, including:

1. *Information literacy and critical thinking skills*: Researchers are typically already highly adept at finding, evaluating, and synthesising information from a wide variety of sources. However, in the context of the additional risks to research integrity presented by generative AI, it would be an ideal time to review the wider support available to researchers for further developing information literacy and critical thinking skills, especially for less experienced researchers, such as HDRs. In addition, developing a set of resources for researchers to discuss these technologies with external research partners (in the community, industry, and government) who may need support in assessing the credibility of information sources, images, and other products of AI tools, would be helpful.

2. *Generative AI literacy skills*: Generative AI techniques, such as LLMs, are sophisticated technologies. While relatively few researchers need to be experts in the technology specifics, any researcher using these tools should at least have a broad understanding of the technology, what it can do and what it cannot. It would be helpful to curate a set of resources to support researchers starting out with generative AI, such as GPT-4 or DALL·E 2, to gain an understanding of the characteristics and limitations of the technologies and their outputs.[21]

3. *Safe discussion spaces and reflective research integrity*: As with any new and potentially disruptive technology, not all the implications of use are immediately evident to its users. Assisting researchers to adapt rapidly but safely may be supported by the creation of safe discussion spaces for open discussion, where researchers can feel confident to share their uses and experiences of AI tools for research and reflect on the principles for research integrity in practice. Sharing best practices and learning from others is critical for effective,

[19] Notwithstanding ongoing discussions on the ethical status and legal implications of AI models trained on copyrighted or unfairly used and unattributed materials of artists and creative practitioners.
[20] See, for example, https://aclanthology.org/W17-4719/
[21] Such as, for example, Prof Karin Verspoor's recent talk "ChatGPT: what it is, what it isn't, and what you need to know" https://youtu.be/gbznDIf13qM

interdisciplinary exploration of both the risks and opportunities of these new tools. Reflective discussions will also prepare researchers for 'moments of integrity' in their day-to-day practice and reduce the likelihood of breaches.

4. *Communities of Practice*: As the use of generative AI matures, it is to be expected that certain types of use that are especially beneficial will emerge, along with associated practices that help researchers more easily conform to and/or expand upon the expected norms and standards in different research fields. Into the future, support for the creation of such research networks and communities of practice of generative AI use may assist in more effectively and safely using the technology as it becomes more deeply embedded in the researcher toolkit.

5. *Adapted services for trustworthy research:* Existing research integrity services and infrastructure might be further enhanced to facilitate responsible research that uses generative AI technology. Increased awareness of ways of conducting responsible research with generative AI through education, engagement, and advisory services will help to support integrity. Appropriate revisions to policy and procedures for authorship and the dissemination of research findings will help clarify expectations for researchers and the University.

Acknowledgements

This White Paper was the result of an open collaboration by academic members of the RMIT *Research Integrity and Generative AI* (RIGAI) Expert Working Group.

Working Group Leads

- Matt Duckham, Enabling Impact Platform Director Information in Society
- Falk Scholer, Professor of Information Access and Retrieval Technologies
- Daniel Barr, Principal Research Integrity Advisor

Working Group Members

- David Blades, Senior Coordinator, RI & Governance
- Chi-Tsun (Ben) Cheng, Associate Professor
- Brian Falzon, Director RMIT Space Industry Hub
- Anthony Forsyth, Professor
- Lisa Given, Professor of Information Sciences & Director, Social Change Enabling Impact Platform
- Dana Mckay, Senior Lecturer Innovative Interactive Technologies
- Anne-Laure Mention, EIP Director Global Business Innovation
- Xavier Mulet, ECP Director, Advanced Materials, Manufacturing & Fabrication
- Magdalena Plebanski, EIP Director Biomedical & Health Innov
- Jason Potts, Distinguished Professor, Economics
- Mark Sanderson, Dean Research & Innovation STEM College
- John Thangarajah, Research Director Centre for Industrial AI Research and Innovat
- Julian Thomas, Distinguished Professor and Director, ARC CoE for Automated Decision-Making and Society
- Karin Verspoor, Dean School of Computing Technologies
- Charlie Xue, Associate DVC, International
- Irene Yarovsky, Professor
- Xinghuo Yu, Vice Chancellors Professorial Fellow
- Fabio Zambetta, Associate Dean Artificial Intelligence