Technology and Regulation

Risk Regulation of Generative AI:

A Case Study of Microsoft Copilot in the Australian Government

Author(s)	Jayson Lamchek and Van-Hau Trieu
Contact	j.lamchek@westernsydney.edu.au, t.trieu@deakin.edu.au
Affiliation(s)	Jayson Lamchek is Lecturer at Western Sydney University, Sydney, Australia. Van-Hau Trieu is Associate Professor in Information Systems at the Business School, Deakin University, Melbourne, Australia.

Keywords	Generative Artificial Intelligence; Information technology; Risk regulation; Practice; Litigation measures; Australian Government; Cybersecurity; AI ethics; Automation bias		
Published	Received: 04 Mar 2025	Accepted: 21 Mar 2025	Published: 14 Jan 2026
Citation	Jayson Lamchek and Van-Hau Trieu, Risk Regulation of Generative AI: A Case Study of Microsoft Copilot in the Australian Government, Technology and Regulation, 2026, 18-37 • 10.71265/esg7cf11 • ISSN: 2666-139X		

Abstract

In many countries, risk regulation is central to AI regulation. We empirically examine risk regulation of generative AI (genAI) through a case study of the trial deployment of Microsoft Copilot in Australian government agencies. Risk mitigation largely depended on end-users' responsibility for human review and fact-checking, readiness testing and contractual assurance from vendors, but largely ignored the impact on team dynamics and long-term implication on human abilities. Improvements to fact-checking and human review by end-users lacking in time, knowledge and experience could strengthen end-user-focused measures. However, impending uses of genAI systems as internal and public-facing government chatbots require other improvements that gesture beyond a 'light touch' approach and the development of government-mandated collaboration among developers, deployers and users.

1. Introduction

Increasing recognition of Artificial Intelligence (AI)'s ethical and social issues[1] triggered states to adopt a conscious approach towards AI regulation. AI regulation may or may not entail enactment of new

[1] Mark Coeckelbergh, 'Ethics of Artificial Intelligence: Some Ethical Issues and Regulatory Challenges' [2019] TechReg 31.

legislations; globally, countries could employ many different kinds and combinations of instruments ranging from legislation to voluntary standards to regulate AI.[2] Whether compliance is obtained mandatorily or voluntarily, however, the instruments that attempt to regulate AI usually seek to require an array of actors, namely, developers, deployers and/or users of AI, to view AI as potentially harmful. These instruments commonly provide for duties to identify, mitigate, monitor and address risks of harm or misuse. Kaminski has called this 'risk regulation' and has theorised duties of this sort as the backbone of much AI regulation.[3]

While there are alternative approaches, risk regulation is the dominant approach to AI in many countries.[4] Partly, this is because of its consistency with the idea that AI's adoption is inevitable. 'By framing the regulation of AI systems as risk regulation, policymakers are, knowingly or not, taking a normative stance on AI. First, risk regulation typically assumes a technology will be adopted despite its harms.'[5] Risk regulation is 'techno-correctionist', i.e., it is geared towards 'try[ing] to fix problems with the technology so it may be used, rather than taking as a starting point that sometimes it might be better not to deploy the technology at all.'[6]

The advent of generative Artificial Intelligence (genAI) technology and the notion of genAI as ushering in unprecedented opportunities has pushed businesses, government organisations, and a host of other actors even more to employ AI to their advantage.[7]

Applying risk regulation to genAI is consistent with narratives of unprecedented benefit or opportunity; it reflects the position that, like AI more generally, genAI as a technology will not, cannot or should not be prohibited though it may be faulty or error-prone; rather, genAI should be made to work properly and effectively to benefit society, organisations or individuals. Yet this focus on enabling opportunity exists alongside the recognition that genAI may generate unintended consequences, underscoring the need for careful governance and risk mitigation.[8]

While viewing the harms of AI systems, including genAI, primarily as systemic risks, instead of say, individualised torts, has certain advantages,[9] adopting risk regulation doesn't mean all risks can be predicted and addressed in advance. Indeed, how risks will arise is often unpredictable. Managing risks as they arise then becomes a necessity especially in the absence of *ex ante* standards. Moreover, whether risks are managed beforehand or in real-time, risk regulation assumes risk managers can 'quantify' risks and proportionately deal with them.[10] Thus, a focus on risk regulation implies that harms that are 'unquantifiable' or not easy to quantify may be missed or remain unaddressed.[11]

Moreover, commitment to a risk regulation approach to regulate AI should not foreclose adoption of features of other regulatory approaches. As Kaminski argues, 'not all risk regulation is the same'.[12] For example, in the European 'risk-based approach' embodied in the EU AI Act, bans are contemplated for

[2] Anu Bradford, *Digital Empires: The Global Battle to Regulate Technology* (OUP 2023); Mona Sloane and Elena Wüllhorst, 'A Systematic Review of Regulatory Strategies and Transparency Mandates in AI Regulation in Europe, the United States, and Canada' (2024) 7 Data & Policy e1.

[3] Margot E Kaminski, 'Regulating the Risks of AI' (2023) 103 Boston University Law Review 1347.

[4] ibid (analysing the text of existing and proposed AI regulation in US and EU); See also, Sloane and Wüllhorst (n 2) (analysing US, EU and Canadian regulations); Mimi Zhou and Lu Zhang, 'Navigating China's Regulatory Approach to Generative Artificial Intelligence and Large Language Models' (2025) 1 Cambridge Forum on AI: Law and Governance e1 (analysing Chinese regulation).

[5] Kaminski (n 3) 1352.

[6] ibid 1354.

[7] To avoid confusion, when we use the term 'AI' we refer to AI in general, which includes generative AI. When only generative AI is referred to, we use the more specific designation 'genAI'.

[8] Hind Benbya, Franz Strich and Van-Hau Trieu, 'Accounting for Unintended Consequences in IS Research: A Call to Action' (2025) 29 Australasian Journal of Information Systems 6063.

[9] See, e.g, Matthew U Scherer, 'Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies' (2016) 29 Harv JL & Tech 353.

[10] Robert Diab, 'Too Dangerous to Deploy? The Challenge Language Models Pose to Regulating AI in Canada and the EU' (1 January 2024) <https://papers.ssrn.com/abstract=4680927> accessed 9 October 2024.

[11] Kaminski (n 3) 1354.

[12] ibid 1369.

certain uses of AI deemed as entailing ‘unacceptable risks’. Bans are posed as consistent with regulation proportionate to the ‘level’ of risk.[13] One reason why an EU AI Act-type regulation is shunned in Australia, however, is the perceived disadvantages of having a comprehensive classification of fixed levels or buckets of risks and regulatory duties. Gikay, for example, argues that the EU AI Act’s risk classification is potentially contentious in practice, and he emphasises the complexity of risk identification and prioritisation.[14] In this light, ‘incrementalism’, as exhibited in the UK’s National AI Strategy and AI Bill, is touted as being ‘more adaptable to evolving risks’ and better able to regulate proportionately.[15] The four characteristics, namely, ‘sectoralism, reliance on existing legal frameworks, evidence-based regulation, and adaptability (flexibility)’[16] that Gikay considers as the hallmarks of the UK’s incremental AI risk regulation can also be seen in Australia.[17]

Australia’s approach hews closely to what Kaminski calls ‘light touch’ risk regulation, i.e., regulation that is ‘focus[ed] on impact assessment and mitigation, much of which is self-supervised and subject to ex post regulatory intervention, if any’.[18] Light touch risk regulation ignores that AI regulators have choices about adopting tools from risk regulation of other domains like the environment or health or approaches outside of risk regulation. For example, it omits ‘civil liability as a backstop to regulatory risk regulation’[19] and other risk regulation tools, namely, ‘precautionary tactics’ like legal bans already mentioned and licensing and ‘post-market measures’ like failsafe modes.[20]

Unlike the European Union or Canada, for example, Australia has neither adopted legislation nor tabled a bill that applies or would apply to entities that develop or deploy AI systems across all sectors. Instead, Australia endorsed AI ethics principles that identify ethical and social issues with AI,[21] and it considers AI systems as subject to technology-neutral Australian laws that apply to all technologies.

In Australia, AI risk regulation within the government sector is exhibited in AI risk management duties imposed on government agencies in accordance with non-legislative instruments called ‘AI assurance frameworks’. Originally championed by the New South Wales (NSW) government in 2022 before the introduction of ChatGPT, AI assurance frameworks govern AI use in government and explicitly cover genAI; a national framework furthermore seeks to align state-level frameworks.[22] In NSW and Western Australia, the assurance framework requires government agencies undertaking an AI project to self-assess their projects against risks defined in relation to five AI ethics principles.[23] They must assess the level of risks from very low or non-applicable to very high, and manage risks accordingly by proceeding as-is, putting mitigation measures in place, or stopping. Thus, potentially, agencies may find that certain measures or mechanisms involving developers and users are required to mitigate and manage AI risks, absent which agencies could

[13] Regine Paul, ‘European Artificial Intelligence “Trusted throughout the World”: Risk-Based Regulation and the Fashioning of a Competitive Common AI Market’ (2024) 18 Regulation & Governance 1065.

[14] Asress Adimi Gikay, ‘Risks, Innovation, and Adaptability in the UK’s Incrementalism versus the European Union’s Comprehensive Artificial Intelligence Regulation’ (2024) 32 International Journal of Law and Information Technology 1, 3.

[15] ibid.

[16] ibid 16.

[17] Christopher T Marsden and Jeannie Marie Paterson, ‘Generative AI Regulation in the UK and Australia: Comparing Two National Attempts at Un-Regulation’, *IET Conference Proceedings* (The Institution of Engineering and Technology 2025).

[18] Kaminski’s own example of this is the self-regulatory approach of the US’ National Institute of Standards and Technology (NIST) AI Risk Management Framework. Kaminski (n 3) 1379.

[19] Kaminski (n3).

[20] ibid 1371–1372.

[21] Department of Industry, Science and Resources, *Australia’s AI Ethics Principles* (2022) https://www.industry.gov.au/publications/australias-artificial-intelligence-ethics-framework/australias-ai-ethics-principles accessed 10 February 2023.

[22] New South Wales Government, *Artificial Intelligence Assurance Framework* (2022) https://www.digital.nsw.gov.au/sites/default/files/2022-09/nsw-government-assurance-framework.pdf accessed 29 July 2024; Department of Premier and Cabinet Office of Digital Government, *WA Government Artificial Intelligence Assurance Framework*; Australian Government, National Framework for the Assurance of Artificial Intelligence in Government https://www.finance.gov.au/sites/default/files/2024-06/National-framework-for-the-assurance-of-AI-in-government.pdf accessed 1 August 2024.

[23] In NSW and WA, larger projects or those with mid-range to higher residual risks, must submit their assessment for review by an AI review body.

decide that the AI project should not proceed altogether. Alternatively, agencies may self-assess their AI projects as requiring no mitigation measures at all and proceed as-is.[24]

In this article, we examine a trial deployment of genAI to support functions within the government sector in Australia. We illustrate how risk regulation is actually being practiced in relation to genAI in our case study sites. Our snapshot of genAI use and regulation in an Australian context shows how government agencies actually try to minimise harm, the challenges they confront and will likely confront as genAI develops further, and therefore the limitations of existing risk regulation. Our case study demonstrates that current risk regulation of genAI applications in government is overly focused on ensuring the responsibility of *users*. While a regulatory focus on users may suffice in relation to low-risk applications,[25] user-focused regulation must also not remain stagnant but must develop as AI applications involve increasing levels of risk. 'Collaboration along the AI value chain'[26] must be developed to more sufficiently handle such applications.

We believe this paper to be one of the first to provide an empirical perspective on how risk regulation of genAI actually plays out. Given the recency of regulatory frameworks for AI and genAI particularly, several analyses of risk regulation have appeared with largely normative or theoretical approaches.[27] Recent studies of AI applications in government are similarly normative or theoretical and do not specifically examine genAI, including a study from a 'design fictions' or future studies perspective;[28] a policy analysis of national AI strategy documents;[29] and literature review of AI-based systems and maintenance of government accountability.[30] We provide a case study of actual deployment of a specific genAI tool within government agencies in Australia, which empirically grounds the exploration we conduct here of the potential and limitations of Australian AI risk regulation as applied to genAI uses that will likely further occur in our research sites. Through our contextual examination, we aim to expose underdeveloped areas of risk mitigation that require urgent attention from Australian risk regulators as new uses for genAI emerge or are proposed that significantly diverge from current uses in terms of risk profile. In making our recommendations in this regard, we are inspired by Kaminski's notion that 'not all risk regulation is the same' and that awareness of regulatory choices could recalibrate existing Australian AI risk regulation.

Part 2 provides a background discussion of genAI, some key risks specific to genAI and their associated mitigation measures. Readers already familiar with the topic may choose to skip this part. We explain our focus on these key risks in the methodology section. Further, these risks are relevant currently and in the immediate future as opposed to risks in the far future and that assume the development of far greater AI capability.[31] Part 3 outlines our research methodology while Part 4 presents findings from interviews.

[24] In NSW, AI projects that simply make use of widely available commercial AI applications which are 'not being customised in any way or being used other than intended' were exempted from compliance with the assurance framework altogether. Moreover, self-assessment of AI projects is required to be undertaken not only as a one-off exercise before project commencement but continuously as risks change throughout the project. AI risk management in government agencies is dependent on and specific to the individual agencies; and agencies have wide discretion about what mitigation measures to require and from whom. Agencies' self-assessment may be affected by factors such as risk managers' knowledge and appreciation of AI risks and mitigation measures, the pre-existing processes of risk regulation in the agency, and agencies' trust in the AI product vendor.

[25] Philipp Hacker, Andreas Engel and Marco Mauer, 'Regulating ChatGPT and Other Large Generative AI Models' in *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (ACM 2023) 1119 https://doi.org/10.1145/3593013.3594067 accessed 28 July 2024.

[26] ibid 1118.

[27] See, e.g., Carsten Orwat and others, 'Normative Challenges of Risk Regulation of Artificial Intelligence' (2024) 18 NanoEthics; Lily Ballot Jones, Julia Thornton and Daswin De Silva, 'Limitations of Risk-Based Artificial Intelligence Regulation: A Structuration Theory Approach' (2025) 5 Discover Artificial Intelligence; Beatriz Botero Arcila, 'AI Liability in Europe: How Does It Complement Risk Regulation and Deal with the Problem of Human Oversight?' (2024) 54 Computer Law & Security Review.

[28] Pedro Vitor Marques Nascimento and others, 'The Future of AI in Government Services and Global Risks: Insights from Design Fictions' (2025) 13 European Journal of Futures Research 1.

[29] Colin van Noordt, Rony Medaglia and Luca Tangi, 'Policy Initiatives for Artificial Intelligence-Enabled Government: An Analysis of National Strategies in Europe' (2025) 40 Public Policy and Administration 215.

[30] Qianli Yuan and Tzuhao Chen, 'Holding AI-Based Systems Accountable in the Public Sector: A Systematic Review' (2025) 48 Public Performance & Management Review 1.

[31] This report does not discuss so-called 'existential' risk ('a scenario in which AI is able to bring about the destruction of humanity') from AI that does not exist yet (also called 'frontier AI'): see Gina Helfrich, 'The Harms of Terminology: Why We Should Reject so-Called "Frontier AI"' [2024] AI and Ethics https://doi.org/10.1007/s43681-024-00438-1 accessed 1 August 2024.

We show how the selected risks were appreciated and managed by Australian government agencies and public officers. From the viewpoint afforded by our case study, Part 5 offers a discussion of how mitigation measures may be further developed. We provide a set of recommendations for the government towards this end while reflecting on how these improvements gesture beyond 'light touch' risk regulation. Part 6 concludes the paper.

2. Theoretical Background

2.1 Generative Artificial Intelligence

Generative AI (genAI) refers to models and systems that create various types of content based on user-supplied prompts.[32] The key to guiding this process is prompt engineering, which is the skill of effectively designing these inputs.[33] GenAI can perform a wide range of tasks, from machine translation and text summarisation to creating original content.[34] GenAI models come in several forms, including large language models (LLMs) for text[35] and specialised models for generating images,[36] as well as audio[37] and video[38]. GenAI's defining characteristic is its ability to swiftly generate diverse and original content during human interaction, distinguishing it from older AI. The user-friendliness and cost-effectiveness of these systems drive their widespread use.[39] Due to how easy it is to use and how widely it's being adopted, genAI significantly changes how organisations create new knowledge[40] and transforms innovation processes.[41] GenAI offers significant opportunities for innovation, efficiency, and decision-making, but it can also produce unintended consequences.[42]

2.2 Risks and Mitigation Measures of GenAI

Scholars and governments alike propose several scenarios where the use of genAI tools based on LLMs could lead to harm to individuals and organisations. The scenarios most recognised by AI researchers include the spread of misinformation, the perpetuation of social inequalities, the leakage of personal data protected by privacy regulation, plagiarism, and the misuse of copyrighted material.[43] Additionally, there are concerns about malicious use (intentional as opposed to unintended harm) of these technologies and their high environmental cost.[44] Mökander et.al. emphasised that each of these scenarios represent a 'complex field of research'.[45] Below we expound on a few of these risks as discussed in the literature and Australian regulations; as explained in Part 3, we focus on these risks in our case study. While only a short list, the selection covers a variety of potential harms and highlights the associated mitigation measures that have

[32] Weng Marc Lim and others, 'Generative AI and the Future of Education: Ragnarök or Reformation? A Paradoxical Perspective from Management Educators' (2023) 21(2) International Journal of Management Education 100790.

[33] Stefano Rizzi and others, 'Conceptual Design of Multidimensional Cubes with LLMs: An Investigation' (2025) 159 Data & Knowledge Engineering 102434.

[34] Sebastian G Bouschery, Vera Blazevic and Frank T Piller, 'Augmenting Human Innovation Teams with Artificial Intelligence: Exploring Transformer-Based Language Models' (2023) 40 Journal of Product Innovation Management 139.

[35] Zenan Chen and Jason Chan, 'Large Language Model in Creative Work: The Role of Collaboration Modality and User Expertise' (2024) 70 Management Science 9101.

[36] Benbya, Strich and Trieu (n 8).

[37] Felix Kreuk and others, 'AudioGen: Textually Guided Audio Generation', (International Conference on Learning Representations, Kigali, May 2023.

[38] Yitong Li and others, 'Video Generation From Text', in *Proceedings of the AAAI Conference on Artificial Intelligence* (2018) 7149.

[39] Leonard Boussioux and others, 'The Crowdless Future? Generative AI and Creative Problem-Solving' (2024) 35 Organization Science 1589.

[40] Maryam Alavi, Dorothy E Leidner and Reza Mousavi, 'A Knowledge Management Perspective of Generative Artificial Intelligence' (2024) 25 Journal of the Association for Information Systems 812.

[41] Sebastian Krakowski, 'Human-AI Agency in the Age of Generative AI' (2025) 35 Information and Organization 100523.

[42] Benbya, Strich and Trieu (n 8).

[43] Emily M Bender and others, 'On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?', in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (ACM 2021) 610; Laura Weidinger and others, 'Ethical and Social Risks of Harm from Language Models' (ArXiv, 8 December 2021) https://arxiv.org/abs/2112.04359 Renee Shelby and others, 'Sociotechnical Harms of Algorithmic Systems: Scoping a Taxonomy for Harm Reduction', in *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society* (ACM 2023) 423; Jakob Mökander and others, 'Auditing Large Language Models: A Three-Layered Approach' (2023) 3 AI and Ethics 361.

[44] ibid, Weidinger and others (n 43); Bender and others (n 43).

[45] ibid, Mökander and others (n 43) 4.

been identified in the literature and Australian regulations. In the below text box, we show the selected risks; the provenance or stage when they arise and the actors (developer, deployer, end-user) involved at that stage; and the associated mitigation measures and the actors involved in executing those measures:

Risks	**Provenance (Actor Involved)**	**Mitigation (Actor Involved)**
misinformation	model pre-training (*developer*)	human review of output (*end-user*)
social inequalities	model pre-training (*developer*)	avoidance of use for decision-making (*end-user*); fine-tuning (*developer, deployer*)
data oversharing within organisations	deployment (*deployer*)	unknown
negative impact on team dynamics	deployment (*deployer*)	license allocation policy (*deployer*)
malicious uses	all stages (*malicious actor*)	cybersecurity (*developer*)

2.2.1 *The Spread of Misinformation*

Large language models (LLMs) have a 'hallucination' problem, i.e., they confidently present claims that are false or completely made up as factual. The issue is not a glitch but an inherent feature of LLMs. As Bender et.al. emphasise, an LLM's performance of language tasks is essentially guesswork where the model generates often-plausible responses (in the sense of well-formed sentences) to a query without complete or real understanding of the question or the content of the response.[46] Therefore, a human user must not always rely on the LLM's output to perform the language task but must make an independent evaluation of the output and act accordingly (in terms of adapting the output before presenting it as a human product). Failing to do this, erroneous, inaccurate or false information may be reproduced, distributed or amplified leading to harms associated with misinformation.[47] The risk arises from 'the processes by which LMs [language models] learn to represent language';[48] while these processes emanate from the pre-training stage, current government-mandated mitigation measures and the responsibility to carry them out are typically addressed to end-users rather than actors who pre-train the models.

The Australian federal government's 'Interim guidance for agencies for government use of generative Artificial Intelligence platforms' (July 2023) (applicable to publicly available generative AI) acknowledges the risk of hallucination. Thus, as a basic guidance, public officers are instructed that '[a]ny responses or outcomes provided by these tools should always be reviewed for appropriateness and accuracy, as they can provide incorrect answers in a confident way.'[49] The interim guidance also recommends flagging the use of AI to inform activities and markings that indicate content generated with AI assistance.[50]

While human fact-checking or review can mitigate the problem, it may not work in all instances. For example, the mistake of trusting an LLM's output can occur for psychological reasons (e.g., confirmation bias), i.e., human users may be sufficiently convinced by the LLM's output and thus dispense with fact-checking. When the LLM outputs correspond with facts with sufficient regularity (though still by chance), human users may develop the habit of trusting in the LLM, heightening the risk of misinformation. Thus, arguably, an LLM that produces outputs that happen to correspond more often with facts may pose a greater hazard for being easier to trust.[51] Further, because of *automation bias*, human users may prefer LLM output to human output for being 'more accurate' or 'more objective' simply because it was automatically generated.

[46] ibid, Bender and others (n 43).

[47] Weidinger et.al. summarise 'misinformation harms' as 'the risk of creating less well-informed users and of eroding trust in shared information.' Weidinger and others (n 43) 1.

[48] ibid.

[49] Digital Transformation Agency, Interim Guidance for Agencies on Government Use of Generative Artificial Intelligence Platforms, Australian Government (2023) 1 https://www.dta.gov.au/help-and-advice/technology-and-procurement/generative-ai/interim-guidance-agencies-government-use-generative-ai-platforms.

[50] ibid 3.

[51] Weidinger and others (n 43) 23.

2.2.2 The Perpetuation of Social Inequalities

The use of AI for automatic decision-making often carries the risk of discrimination against vulnerable groups; this is true for genAI as well. In Australia, the concern with algorithmic bias and automated decision-making facilitating unaccountable decisions has been acknowledged prior to ChatGPT.[52] Biases arise from the characteristics of the datasets in which the model was trained. Training data may be unrepresentative or contain statistical patterns or reflect social values at a point in time; these patterns and values are then coded within AI models. Thus, AI could reproduce derogatory associations and outdated social values (racism, sexism, etc.), perpetuating and amplifying social inequalities and inhibiting social change. Furthermore, affected individuals may not know about those decisions or how the AI model arrived at them or the datasets on which the AI model was trained and therefore may not be able to contest those decisions.[53] Some proposed mitigation strategies relate to the origin of the risk, namely, curating training datasets to increase their diversity or fairness, and documenting training datasets (e.g., through datasheets) to enable the tracking of the training data that affected the AI's predictions.[54] However, it is unclear how these strategies apply to genAI.

Because LLMs are designed to mirror natural language, the risk of reproducing bias embodied in natural language is heightened in LLMs. For example, ChatGPT's training dataset includes content scraped from the entire internet, including all sorts of biases contained therein. It has been demonstrated, for example, that ChatGPT has learned racist, sexist and otherwise toxic language, ideas and opinions from being exposed to the same during pre-training.[55] It also favours certain languages, cultures or perspectives because material reflecting those languages, cultures or perspectives are more available on the internet.[56] The datasets on which LLMs have been pre-trained have not been curated. Furthermore, it may be practically impossible to document the data that affected the particular outputs of LLMs, preventing affected persons from understanding the decisions made with LLMs' assistance. When LLMs have these characteristics, LLMs are essentially black boxes that are not advisable for use in decision-making settings. Quite apart from their use in decision-making, exposure by humans to LLM bias may be harmful in itself as humans may replicate bias beyond their interaction with LLMs.[57]

To prevent harm of this nature, the Australian federal government's end-user guidance alluded to earlier cautions against using freely available generative AI (such as ChatGPT) in 'high-risk situations'. It considers 'use cases where services will be delivered, or decisions will be made' as 'use cases which currently pose an unacceptable risk to the government'. It reiterates that '[g]enerative AI tools must not be the final decision-maker on government advice or services. Accountability is a core principle for activities within the APS [Australian Public Service]. As such, humans should remain as the final decision maker in government processes.'[58]

[52] See, e.g., Sophie Farthing and others, Human Rights and Technology Final Report 2021 (Australian Human Rights Commission 2021).

[53] Cathy O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (Crown 2016); Virginia Eubanks, *Automating Inequality* (St Martins Press 2018).

[54] Bender and others (n 43) 615.

[55] Samuel Gehman and others, 'RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models' (arXiv, 25 September 2020) http://arxiv.org/abs/2009.11462 accessed 29 July 2024; Li Lucy and David Bamman, 'Gender and Representation Bias in GPT-3 Generated Stories' in Nader Akoury and others (eds), *Proceedings of the Third Workshop on Narrative Understanding* (ACL 2021) 48; Abubakar Abid, Maheen Farooqi and James Zou, 'Persistent Anti-Muslim Bias in Large Language Models' (arXiv, 18 January 2021) http://arxiv.org/abs/2101.05783 accessed 29 July 2024.

[56] Xavier Ferrer and others, 'Discovering and Categorising Language Biases in Reddit' (arXiv, 13 August 2020) http://arxiv.org/abs/2008.02754 accessed 29 July 2024; Eun Seo Jo and Timnit Gebru, 'Lessons from Archives: Strategies for Collecting Sociocultural Data in Machine Learning', in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (ACM 2020) 306.

[57] Lauren Leffer, 'Humans Absorb Bias from AI—And Keep It after They Stop Using the Algorithm' (2023) Scientific American https://www.scientificamerican.com/article/humans-absorb-bias-from-ai-and-keep-it-after-they-stop-using-the-algorithm/.

[58] Australian Government Digital Transformation Agency (n 49).

While bias is acquired during model pre-training, there are proposed strategies to manage the risk of bias in LLM after the pre-training stage. For example, during the fine-tuning stage, the LLM's response can be subjected to a test for appropriateness and then moderated accordingly before being sent to the user.[59]

2.2.3 Information Oversharing within Organisations and Impact on Team Dynamics

Focusing on LLM deployed at the workplace, management literature raises a different set of concerns, namely, information oversharing within organisations and the impact on team dynamics. Microsoft's Copilot for Microsoft 365 is an example of an AI system which integrates LLMs with a client organisation's internal data. The Copilot system consists of three components: 'Microsoft 365 apps such as Word, Excel, and Teams, where users interact with the AI assistant; Microsoft Graph, which includes files, documents, and data across the Microsoft 365 environment; and the OpenAI models that process user prompts: OpenAI's ChatGPT-3, ChatGPT-4, DALL-E, Codex, and Embeddings. These models are all hosted on Microsoft's Azure cloud environment.' According to Finnegan, the main data security concern with Copilot isn't primarily about sharing data with Microsoft but rather the risk of internal information exposure.[60] IT departments who are traditionally tasked with securing sensitive documents must ensure that the introduction of Copilot does not increase the likelihood of data breaches. Although Copilot's access to files is limited to the permissions assigned to each employee, the issue arises when companies do not properly classify confidential documents. Consequently, employees may discover that they can request Copilot to retrieve information on payroll or customer data if these documents are not secured with appropriate permissions. How organisations are addressing this challenge is not well-known.

While the deployment of Copilot could help employees enhance their productivity and efficiency, the distribution of licenses on a selective basis within teams could lead to negative outcomes. Cooper highlighted the potential ethical issues of Copilot usage in terms of imbalanced competitive environment, impact on pay and career advancement, inclusion, equality, and diversity, and team morale.[61] Thus, selective allocation of Copilot licenses may mean certain individuals may be provided a considerable advantage over co-workers, potentially resulting in more favourable performance reviews, increased salary, or accelerated career advancement. This potentially raises questions around fairness that could generate conflict and disgruntlement among employees. Besides fostering feelings of inequality, the organisation's values and efforts towards diversity and inclusion could be undermined. Broadly, Cooper suggests that managers should have a well thought out policy on allocating licenses as well as measures that engage seriously with these ethical issues.

2.2.4 Malicious Uses and Cyber Security of LLMs

Besides the harmful unintended consequences identified above, LLMs could also be used intentionally to cause harm. Weidinger et.al. identify some possible examples of intentional or malicious use of LLMs by or affecting governments.[62] First, governments themselves could use LLMs to automate mass surveillance of speech/texts which would have a negative impact on freedom of expression. Second, malicious actors could try to use LLMs to frustrate government processes, e.g., flooding the government with fake submissions.[63] Disinformation campaigns may become more effective with more powerful LLMs. Examples of concern to Australian regulators are the Australian Competition and Consumer Commission (ACCC) concern with fake product reviews, and the eSafety Commission's concern for deep fakes (sexual abuse images especially targeting children).[64]

[59] Australia's Chief Scientist, Generative AI: Language Models and Multimodal Foundation Models (Rapid Response Information Report 2023) 4–5.

[60] Matthew Finnegan, 'M365 Copilot, Microsoft's Generative AI Tool, Explained' (*Computerworld Australia*, 1 November 2023) <https://www.computerworld.com/article/3700709/m365-copilot-microsofts-generative-ai-tool-explained.html> accessed 20 March 2024.

[61] Sam Cooper, 'Navigating the Ethics of Assigning Microsoft Copilot Licences' (*Changing Social*, 12 September 2023) https://www.changingsocial.com/blog/copilot-ethics/ accessed 2 February 2024.

[62] Weidinger and others (n 43).

[63] Digital Platform Regulators Forum Joint Submission to Department of Industry, Science and Resources – Safe and Responsible AI in Australia Discussion Paper (Digital Platform Regulators Forum 2023) para 4.2.

[64] O'Loughlin and others (n 63).

The risk of malicious use is related to the cybersecurity of LLMs themselves. As is the case with traditional or mainstream digital technologies, platforms and software, LLMs are vulnerable to attacks by malicious actors seeking to diminish model integrity, accuracy or availability.[65] Both the training dataset *and* the machine learning (ML) model (as intellectual property) are targets for malicious actors. The literature suggests that known attack types can be broadly separated either by the stage targeted (in the phase of ML training; or during the test and application process) or by the objective of the attack itself (attacking the integrity of an LLM; or compromising a model's confidentiality). Because malicious actors will act maliciously regardless of any rules or guidance put in place to prevent misuse of LLMs, addressing risk from malicious uses tend to go beyond a focus on end-users and into the secure design and development of LLMs themselves.

3. Methodology: a Case Study Approach

3.1 Case Study: Sites, Risk Selection, and Limitations

In 2023, the Australian government announced that it was innovating with AI by introducing the use of genAI tools to enhance business efficiency and support public officers in the exercise of their functions.[66] The sites for our case study were two large Australian government units (each consisting of several agencies) in which pilot or trial deployments of the genAI tool Microsoft Copilot were conducted. In accordance with the Australian government's AI policy, it acknowledged that AI innovation must be done safely and responsibly, i.e., organisations must not only discover new beneficial uses for AI but must do so while minimising the risks of harm.[67] We examined the risk management of Microsoft Copilot performed by the relevant officials as provided in relevant AI policy and regulations. Our respondents included not only end-users but also deployers (with organisational risk management duties with purview over AI merchants and developers) and policy makers (with purview over AI merchants and developers).

At the time of interview, agencies were running pilot or trial phases of Copilot, involving only a select group of public officers within each agency before full deployment. Prior to the trial run itself, agencies commonly conducted readiness testing to ensure the smooth technical integration of Copilot with their information systems. As respondents involved in the product deployment explained, Copilot is given access to the organisation's internal documents or sites so that it could answer queries based on those documents or sites. Individual users access the licensed version of Copilot from the Microsoft applications such as Teams, Outlook, Excel and PowerPoint. Copilot is enabled to show specific users all and only the internal documents that the specific user has permission to access. The pilot stage also involved the conduct of Microsoft-designed end-user training for the initial users which included guidance about ethical use. Through pilot deployment, agencies established the business case for full deployment, ascertaining the use cases for Copilot that produced value for agencies as well as considering the associated risks and the needed mitigation measures going forward.[68]

To explore the practice of risk regulation of genAI in the Australian government, we supplemented search and textual analysis of Australian regulations with discussion of the selected risks expounded in Section 2 with research participants.[69] The selection of these risks was influenced by specific circumstances of the Australian context and case study. The misinformation and social inequality risks were included because

[65] See, generally, Andrew Lohn, 'Hacking AI: A Primer for Policymakers on Machine Learning Cybersecurity' (Center for Security and Emerging Technology 2020) https://cset.georgetown.edu/publication/hacking-ai/ accessed 8 December 2022.

[66] Anthony Albanese, 'Australian Government Collaboration with Microsoft on Artificial Intelligence' (Media Release, 16 November 2023) https://www.pm.gov.au/media/australian-government-collaboration-microsoft-artificial-intelligence accessed 29 July 2024; Digital Transformation Agency, 'Australian Government Trial of Microsoft 365 Copilot: Summary Report' (n.d.) https://www.digital.gov.au/initiatives/copilot-trial/summary-evaluation-findings/cts-executive-summary accessed 29 July 2024.

[67] In Australia, Responsible Artificial Intelligence has been defined as 'the practice of developing and using AI systems in a way that benefits individuals, groups, and the wider society, while minimizing the risk of negative consequences.' CSIRO, 'Responsible AI Pattern Catalogue' (28 September 2023) https://www.csiro.au/en/research/technology-space/ai/Responsible-AI/RAI-Pattern-Catalogue accessed 29 July 2024.

[68] Albanese, (n 66).

[69] We searched for all Australian AI regulations whether AI end-user focused, AI deployment focused and AI design and development focused.

the Australian government itself recognised them as of high importance. This recognition is affirmed in early public pronouncements and interim regulation on genAI risks (referenced in Section 2). Similarly, the inclusion of cybersecurity risk of LLMs used in government is justified by the prominent concern with data leakage (further expounded below) expressed by government agencies. In addition to these, information oversharing and team dynamics risks were included to explore specific organisational and employee concerns arising from the trial deployment.

The adoption of the licensed version of Copilot was shaped by concerns about public officers using freely available genAI tools. The free or public version of ChatGPT was launched in November 2022. This was followed by the release of other genAI tools and the integration of genAI into browsers like Microsoft's Bing and other applications like Meta's Facebook Messenger. As some respondents explained, many public officers experimented with ChatGPT and other public genAI tools to explore their uses for their professional and personal interests. However, the use of ChatGPT by public officers to support their work threatened the confidentiality of government information. This was because data shared by end-users through prompts are stored and used by OpenAI to further train its LLMs thereby potentially exposing those data to the risk of leakage or attacks. Because of this main concern, some agencies like the Department of Home Affairs banned its employees from using ChatGPT altogether.[70] However, most agencies opted instead to educate public officers about the risks associated with using freely available generative AI tools and the proper use of these tools. In this regard, agencies issued basic end-user guidelines that require public officers not to input confidential or sensitive information into prompts.[71]

While the licensed version of Copilot and the latest version of ChatGPT are similar in capability in that both are based on the same pre-trained models GPT3 and GPT4, Copilot was supposed to avoid the concern with data leakage. Microsoft guaranteed to securely store and process the client organisation's data within the latter's tenancy and Microsoft wouldn't use them for further training of its pre-trained model unless users allowed it.[72] Relying on this assurance, Copilot could be provided access to a client organisation's internal databases which weren't publicly accessible on the internet. Thus, Copilot was touted as enabling safer experimentation with genAI use compared with free genAI tools. Public officers could input information they weren't supposed to share with free genAI, applying genAI capability on agencies' internal data.

We acknowledge that the selection of risks discussed with participants represents a limitation of the study. Our findings only include recommendations related to improvement of regulation relating to the selected risks and risks discussed by participants. They do not imply that other risks are less important nor do we make assertions about prioritisation of one set of risks over another. Our intention is simply to utilise our findings to indicate how current Australian government genAI risk regulation should change.

3.2 Data Collection and Analysis

We conducted semi-structured interviews in May and June 2024 with a total of 14 key participants who were directly involved in the Copilot implementation at two major sites. To ensure representation of diverse perspectives within the Copilot implementation process, participants were selected based on their roles in the deployment, use and governance of GenAI tools. Nine participants were drawn from Site A and five from Site B. Thirteen were end-users of Copilot, ten held organisational or product risk management roles other than as end-users, and four were responsible for defining AI-related policies. This combination captured the perspectives of users, deployers, and policymakers to provide a comprehensive understanding of Copilot implementation and governance processes.

The interviews, which lasted between 40 and 60 minutes, were audio-recorded, coded, and analysed. We applied a flexible pattern-matching approach, starting with a coding template informed by themes identified

[70] Ry Crozier, 'Home Affairs Blocks Public Servants from Using ChatGPT' (iTnews, 23 May 2023) https://www.itnews.com.au/news/home-affairs-blocks-public-servants-from-using-chatgpt-596130 accessed 29 July 2024.

[71] See. e.g., Australian Government Digital Transformation Agency (n 49).

[72] Mechanics Team, 'How Microsoft 365 Copilot Works' (*Medium*, 23 May 2023) https://officegarageitpro.medium.com/how-microsoft-365-copilot-works-f3f46f98c9ff accessed 29 July 2024.

in the literature while remaining open to new codes emerging from the data. These codes were then applied to the interview transcripts to identify and analyse themes.[73]

4. Findings

Copilot is an off-the-shelf vendor product that risk managers could not take apart to check for inherent risks. Neither could risk managers address the product's risks by modifying the product. Rather, Copilot risk management is focused on risks that arise from actual use of the product and users are primarily made responsible for mitigating against those risks. As a result, risk management of Copilot was *end-user-focused*.

Respondents uniformly emphasised that Copilot was intended as a work productivity enhancement tool. Copilot's assistance to public officers in the execution of their day-to-day tasks would make writing emails, summarising meetings and reports, generating presentation slides, etc. quicker to accomplish and less of a drudgery. However, Copilot's capability was limited, and active human direction of Copilot was required to ensure it created value. Thus, agencies regarded proper use as limited or controlled use, or as many respondents put it, 'as Copilot not auto-pilot'. It was understood that if not properly limited or controlled, Copilot can pose risks.

4.1 Risk of Spreading Misinformation

Respondents who have used genAI tools were all aware of the hallucination problem of LLMs. They have personally encountered instances when ChatGPT has presented completely made up or false information in a confident manner as if they were facts. Though hallucination is inherent in LLMs and can never be ruled out, many respondents believed that the tendency to hallucinate was probably lessened in the licensed version of Copilot because Copilot has access to the organisation's internal documents and sites. Therefore, its responses were trained on more relevant information compared to other genAI tools that did not have access to internal documents and depended on information from the internet.

Respondents said Copilot's referencing feature helped to some extent in detecting irrelevant or false information in generated responses. Copilot's responses incorporated references to internal documents and sites, the links to which are provided at the end of the response (specific users may or may not be able to access their content depending on their permissions). Thus, Copilot users are enabled to quickly fact check or review the response against the referenced documents or sites (assuming users could access the contents of those documents). Some respondents said by checking the references, they can ascertain, for example, that the response was based on outdated documents, and they therefore alert themselves to adjust or modify the generated content before using it if at all.

Respondents also mentioned the role of effective prompts in increasing the likelihood of generating more relevant responses. Agencies are hoping that teaching users to create well-crafted prompts would help elicit responses from Copilot that are useful for users. Some suggested prompts are already embedded in dropdown menus on Copilot; while prompting 'skills' are currently provided as part of Microsoft's user training and further developed through 'communities of practice' established among users within agencies that share prompting tips with each other.

In practice, as some respondents noted, the value of Copilot as a productivity enhancement tool is affected by how much time and effort users have to put into fact-checking or reviewing its responses. The expectation of agencies is that after users have fact-checked and reviewed generated content, users would still have net gain in terms of time saved on tasks than if public officers did not use Copilot's assistance at all. However,

[73] David Silverman, *Interpreting Qualitative Data* (4th edn, SAGE 2011); N Sinkovics, 'Pattern Matching in Qualitative Analysis' in Catherine Cassell, Ann L Cunliffe and Gina Grandy (eds), *The Sage Handbook of Qualitative Business and Management Research Methods* (SAGE 2018); Robert K Yin, *Case Study Research: Design and Methods* (3rd edn, SAGE 2003); Van-Hau Trieu, A Burton-Jones and S Cockcroft, 'Applying and Extending the Theory of Effective Use in a Business Intelligence Context' (2022) 46 MIS Quarterly 645.

some respondents said that for some tasks, 'it was much more of a hassle' to use Copilot because its response took too much time to modify to be useful.

One respondent who used LLMs to assist with technical problem-solving said Copilot could mislead users who lack sufficient ability to verify the response. This meant that Copilot could, contrary to the intention, create inefficiencies rather than enhance productivity in certain instances. LLMs could generate fabricated yet seemingly convincing solutions to technical queries. An inexperienced person could 'waste a lot of time' using Copilot's response to such a query even merely as a lead or suggestion that needs further verification or investigation.

Most respondents said they considered the risk of spreading misinformation to be 'low' and 'easily manageable'. This was because they used Copilot for tasks that did not involve facts that couldn't easily be verified by the user. For example, Copilot was commonly used to summarise documents or meetings. In these instances, the risk could take the form of hallucination issues, for example, of Copilot misstating salient points in documents or misrepresenting that certain discussion points were raised by certain participants in meetings when they did not. These mistakes can easily be reviewed by the user by referring to the documents or the meeting recordings themselves if needed. The users themselves may be an expert on or sufficiently familiar with the topic of the documents or meetings to detect that mistakes have been made in the response. However, lengthy documents or meeting recordings could still be laborious to review. Hence, one respondent mentioned that he wished Copilot further supported users to look up where exactly in the document or meeting recording the point being summarised appears.

Another use for LLMs that many respondents valued was for generating ideas. One respondent explained that LLMs help him to 'think outside his own brain', that is to generate ideas that may not be on top of his head because of his particular background or interests. However, using an LLM for information search, particularly information search on the internet, entails a different level of risk as it calls the effectiveness of human oversight of an LLM's output into question. Traditional information search involved entering key words into Google (or, in case of internal documents search, the relevant internal search engine); reading through the list of sites returned; judging which sites are more or less relevant; sifting through the individual sites for the relevant information. Each of these steps afforded some degree of human judgment. However, with a genAI tool, these intermediary steps are eliminated. Instead, there is only a single opportunity to fact check and review the generated output which in effect replaces the exercise of human judgment which was previously dispersed among the intermediary steps. Thus, there is a greater risk that final human oversight will fail to effectively correct the errors.

Some respondents conceded that in extreme situations, such as in the presence of productivity or time pressures, a user might potentially turn to Copilot to fully automate certain tasks, dispensing with the required fact-check or review. For example, one respondent suggested, public officers who are overly burdened with replying daily to hundreds of emails from the public might problematically rely on Copilot to auto-generate and send replies without review to cope with the volume of work. This would be seen as a misuse of Copilot.

However, there is a demand to augment human workers through automation in this regard. Several respondents foresee leveraging the capability of LLMs seemingly 'to understand human language' to augment government's capacity to process voluminous public-facing services, for example, as a government chatbot that dealt with queries from the public. This future use of LLM will entail a reassessment of the risk of spreading misinformation as the potential harm might then be borne by the public.

Respondents' observations and views on AI content flagging or marking were also revealing. While AI content marking is suggested in user guidance documents, it is not yet widely practiced and there seems to be a lack of clarity on what the guidance amounts to. Many respondents highlighted that they have not seen important work products, such as briefing notes which contain a by-line, that are genAI content-flagged/marked. Respondents offered various explanations for this. Some consider uses of genAI, such as for wordsmithing, too inconsequential to require that the use of genAI be revealed. Some respondents said that if a genAI-generated content is fact-checked, reviewed or substantially modified by a human (as they are supposed to do), then they considered AI content-marking of the final product unnecessary beyond a general statement that AI assistance was used in its production. This was consistent with the guidance that the public officer must take responsibility for his work product whether or not AI assistance was used. Another respondent mentioned that he used Copilot to expand a shorter speech or calibrate its tone. He considered the output as based on his own work and implied that citing Copilot as co-author in this case would be unnecessary. A respondent emphasised that a work product that is genAI content-flagged/marked simply conveyed the unwanted impression that it was of a lower quality. '*It's not a good look – this perception that you used an LLM to do this work for you*.'

4.2 Risk of Perpetuating Social Inequalities

Most respondents also considered the risk of perpetuating social inequalities to be 'low' for the intended uses of Copilot as a productivity enhancement tool. While conceding the risk exists, respondents said they have not encountered LLMs exhibiting clear examples of bias and prejudices or responses from a discriminatory perspective in the context of their use of Copilot.

The risk of perpetuating social inequalities is particularly relevant to the use of AI for decision-making. In the AI assurance frameworks, agencies are aided to consider this risk in relation to AI ethics principle of fairness and transparency. As black boxes, when used for decision-making, LLMs challenge administrative law principles and requirements regarding the explainability of decisions and the possibility of appeals. Nevertheless, when LLMs are used for other purposes, such weighty considerations become largely irrelevant. A respondent explains: '*My understanding is that that's really for government officers to be aware of risks in ADR [automated decision-making] and to mitigate against bias, job loss, etc. which aren't the biggest risks involved in Copilot.*'

At one level, there is a difference between the intended uses of Copilot for such tasks as summarising documents and meetings and drafting emails and presentation slides, on the one hand, and the use of AI for analytics that supported government decision-making. As one respondent emphasised, agencies are "*not giving Copilot massive datasets and then asking it 'what is the best social policy based on all the documents we've produced over the last four years?'*" Instead, a public officer is inputting a document into Copilot and asking what the salient points in this document are. While the risk of harm from AI perpetuating historical bias is serious in the former, it was largely irrelevant in the latter.

On another level, the distinction between day-to-day tasks and decision-making is not so clear as day-to-day tasks may include activities adjacent to decision-making. Products of day-to-day tasks may feed into decision-making or may themselves constitute decisions of some kind. For example, a well-known context in which automated decision making has been discussed is hiring or recruitment.[74] One respondent said an experimental use of Copilot in his agency was a 'first pass' on a job applicant's CV which checked how much a CV responded to the selection criteria. While this was a use of Copilot to support day-to-day tasks (analysing a document), it also clearly fed into a decision-making process that could affect individual third parties.

It may be argued, in any case, that the risk is sufficiently mitigated by users being conscientious about not substituting Copilot's assistance for their own exercise of judgment. However, similar to using LLMs to automate information searches, careful consideration must be given to the effectiveness of human oversight over an LLM's output when they are used to automate execution of actions that previously required human judgment. Agencies could leverage LLMs to automate and mechanise various labor-intensive processes beyond the day-to-day tasks we have so far discussed. One potential use of LLMs is to automate workflows by enabling workflows to be triggered through conversational input rather than the traditional paper or online forms with the associated administrative processes and controls. One respondent pointed out, for example, that it is possible to integrate an LLM into the system for lodging requests with Human Resources.

[74] See, e.g., Jeffrey Dastin, 'Insight - Amazon Scraps Secret AI Recruiting Tool That Showed Bias against Women' (Reuters 11 October 2018) https://www.reuters.com/article/idUSKCN1MK0AG accessed 12 August 2024.

There will be an even further escalation of risk in case LLMs are deployed as an interface between government and the public in the form of government chatbots.

For example, one respondent explained, an LLM could be used as a government chatbot that provided advice to customers on how to proceed with lodging a request with the government. If it misled people, '*then (A) it opens up the government to being sued, and (B) to having negative outcomes applied to people in the community*.'

4.3 Risk of Oversharing Information

As we have mentioned, the concern for the security and confidentiality of government data in relation to the free version of ChatGPT motivated agencies' adoption of the licensed version of Copilot. While Microsoft assures the safety of sharing organisational data with Copilot in that no data should leak outside of the organisation, a different risk involves oversharing of information within the organisation. Respondents who managed the Copilot trial run said this was a concern for the government; the safe integration of Copilot must ensure that employees could access all and only the internal documents to which they had the appropriate permissions.

However, prior to Copilot, because of human error or a lax attitude, some sensitive documents and sites have been created without the required security. The readiness testing for Copilot would have identified many of these documents and rectified the error but some such documents may still remain. Hence, links to documents that shouldn't be accessible to a user may still show up in the references in Copilot's responses. While this was not a problem created by Copilot, Copilot may sometimes make employees aware of documents they shouldn't have access to and enable employees without the required permission to easily access them. In this sense, the ease of document search afforded by Copilot was contributing to the realisation of a preexisting risk. Respondents are managing this risk by informing users that this was a possibility and by establishing a procedure for reporting and correcting the information security breach.

A respondent warned that the risk of oversharing sensitive information to the genAI tool will be heightened when an LLM is deployed as a government chatbot that interfaces with the public. People may overshare personal information when using a government chatbot because of the public's trusted relationship with the government. For example, a user of a government chatbot may unnecessarily share her tax file number in the hope of obtaining a personalised response to a query.

4.4 Impact on Team Dynamics

The impact of Copilot and other LLMs on public officers' individual work productivity is uneven. This is partly because not everyone at the moment has a Copilot license; and partly because the interest and ability to use Copilot to one's advantage is unevenly developed. In turn, this situation can create a sense of imbalance or tension within organisations. Respondents highlighted the need to have a well-articulated basis for distributing access to AI capability, the importance of developing AI skills and knowledge, and equity considerations.

They acknowledged that there was an existing or potential new digital divide among those with the ability to use AI assistance to one's advantage and those lacking in such ability. Early adopters of genAI tools are seen to have an advantage as they develop beneficial AI skills and knowledge. Some doubted whether the productivity advantage from the current Copilot was substantial enough to have an impact on users' chances of moving up the ranks versus non-users. Others, however, think the ability to use AI to one's advantage was so important that it will have a similar impact on career advancement as one's social connections or having the right mentors.

Some emphasised the positive impact on equity. For example, a respondent noted that translating between English and certain foreign languages via LLMs potentially meant employees whose first language is not English could use their native language at work. Another respondent said he believed Copilot mostly benefited those who are less naturally skilled at drafting emails, etc. and was therefore contributing positively to giving employees an equal playing field.

4.5 Malicious Uses and Cyber Security

Significantly, in none of our interviews did respondents refer or discuss mitigation measures addressed to other actors in the AI value chain besides the end-users. However, some respondents hinted at the limits of user responsibility in preventing or mitigating harm particularly in relation to the risk of harm from malicious uses of LLMs. They affirmed that they were aware of both actual and theoretical vulnerabilities of the models themselves to cyber attacks including attacks before deployment. One respondent expressed concern that the government has no knowledge of what data proprietary LLMs were trained on and how vulnerable they might be to manipulation. Yet it was not clear that Microsoft needed to demonstrate certain security standards before Copilot was deployed.

Malicious uses also bring into focus further pressures on the government to use AI to counter malicious uses and thus to assimilate AI capabilities including LLMs into cybersecurity functions. In turn, this introduces similar risks we have seen in the context of day-to-day tasks of more policy-oriented roles into those associated with cybersecurity roles.

Many respondents expressed awareness that malicious actors can leverage LLM capability to automate and increase the sophistication of attacks on government. A respondent mentioned that a simple example of a malicious use that could harm the government is to barrage the government with fake queries from the public. As the respondent explained, responding to a single query from the public can entail a cost of up to thousands of dollars in government resources. Cleverly constructed queries that were difficult to dismiss as inauthentic could still require the government to respond and, thus, a barrage of such fake queries could potentially paralyse the government.

In reaction, agencies were eager to leverage genAI to better defend against attacks. Already, some agencies are exploring the development of LLMs or utilising existing LLM tools like Copilot for Security for use in cybersecurity contexts. (While Copilot for Microsoft 365 works with Microsoft applications we have so far alluded to, Copilot for Security works with Microsoft applications like Sentinel and Defender which are used to support cybersecurity functions.) In the case of a cybersecurity project supported by one agency, an LLM was used to assist junior analysts by providing them with 'worded up' analyses of technical information. According to the public officer who risk-assessed this project, the risk that the LLM would produce erroneous conclusions was mitigated by ensuring that the LLM 'does not make the final decision', i.e., that a human will still check the 'worded up' analysis.

4.6 Long-term Implication on the Exercise of Valued Human Abilities

Apart from the risks discussed above, many respondents expressed concern for the long-term consequences of integrating genAI assistance at work. In the long term, as genAI assistance simply becomes a natural part of working, some respondents say this could lead to dependence on AI and loss of certain human cognitive abilities or qualities arising from lack of practice. This potential problem also depends on whether end-users succeed in utilising genAI tools to increase the quality of their work or simply to get by with the demands of work.

One respondent explained that the 'auto-reply functionality in Copilot which could through time adapt to your style' tended to eliminate the task of writing which is 'a valuable way to shape your thoughts'. To this respondent, therefore, automating writing tended to take away from being thoughtful. The respondent also suggested that there could also be a loss of excellence or creativity given that LLMs are meant only to produce plausibly human content as opposed to true insight or novelty. 'LLMs just give you the most expected, obvious things that make you lose sight of all the nuance and interesting other analyses.' Echoing this concern, another respondent said he was worried about LLMs facilitating the proliferation of 'slop' as opposed to polished work. To him, LLMs make it easy to generate content that could '*just be slightly wrong*' resulting in a morass of work products that was harder to '*wade through*'. Thus, ironically, really valuable information could be harder to find in the long term.

In a more optimistic view, there could be a mutuality between the increased use of genAI and the enhancement of human abilities. Some respondents say that LLM assistance wouldn't lead to loss of critical thinking skills assuming everyone was responsible in their use of genAI tools. Another respondent summarised the needed cognitive ability to function in a new setting where genAI was '*a normal part of how we function in society*' in terms of self-awareness. 'The way to mitigate against manipulation through AI is to increase awareness of when we are dealing with AI; what information you're providing and what information you're receiving; and your reliance on that information.' These views suggested that there should be intervention, including education and training, intended for optimising outcomes for human abilities in the long-term.

5. Discussion

This section discusses our findings with a view to distilling recommendations that improve on the observed shortcomings of mitigation measures already in place. We depart from a purely empirical investigation to reflect on what these improvements could mean for 'light touch' risk regulation of genAI in the Australian context.

We have seen that risk regulation enables Australian government agencies to consider the application of Copilot and other genAI tools, despite their inherent risks, as a responsible practice. They have assessed risks in the application of genAI tools to support day-to-day tasks of public officers as manageable through end-user-focused mitigation measures. Specifically, the risks of spreading misinformation and perpetuating social inequalities are confidently assessed as either low or irrelevant because risk managers trust public officers to fact check and review AI-generated content and not substitute AI assistance for their own exercise of judgment.

However, this confidence in existing end-user guidance can be challenged. First, end-user-focused mitigation measures, particularly fact-checking and human review can fail. We outline improvements that could strengthen fact-checking and human review processes, particularly for users who may lack sufficient time, knowledge, or experience.

Second, even use cases that pose a low risk in the short term can have important impacts on team dynamics and other implications in the long-term. And lastly, impending applications of LLMs will require elaboration of developer and deployer responsibilities and collaboration on risk management and accountability across the AI value chain. These last two areas of possible improvements to genAI risk regulation could push Australia beyond a 'light touch' version of risk regulation.

5.1 Improvements to User-Focused Mitigation Measures

Fact checking and review can require time and the requisite skill or knowledge to accomplish effectively depending on the task. Time-poor, inexperienced or unskilled users can easily slip into irresponsible usage. For example, searching information which the user, for lack of knowledge, could not verify or would have difficulty verifying is an unsafe or inappropriate use of LLMs. Similarly, a user may, for lack of skill or time, fail to improve upon a generated response which needs revision. As LLMs become more accurate, reliable and sophisticated, user training and education to develop public officers' AI skills and knowledge will become even more rather than less necessary. This is because a more capable LLM can mislead humans more and make them more reliant on LLM assistance.

Time is important for the exercise of judgment and critical thinking. Introducing LLM assistance to justify increasing the volume of work of already overworked public officers would undermine the requirement to exercise judgment and critical thinking when using genAI tools. Care must be taken to ensure that LLMs actually enable workers to perform better at their tasks in a qualitative rather than merely quantitative sense. Moreover, experiential learning and collaboration among users may become important sources of AI skill and knowledge for users. 'Communities of practice' are currently established within agencies for users to share learnings, such as novel use cases and effective prompts. Users learn as they use the genAI tool; training based on theories and abstract principles will not cover all the bases.

Existing user guidance on AI content flagging and marking must be clarified. For example, in Western Australia, the user guidance simply states, '*Where required, attribute content that has resulted from the use of these tools.*'[75] It is not clear when it is required to attribute content to the genAI tool. The Commonwealth guidance is as follows:

> It should also be clear when AI tools are being used by the government to inform activities. Users could consider including markings in briefings and official communications indicating if AI was used to generate any of the information.

The guidance suggests that the goal of 'markings in briefings and official communications indicating if AI was used in *any* of the information' is to reveal when the government uses AI tools to 'inform activities'. However, public officers may not be convinced that it is appropriate or realistic to expect them to attribute content to the genAI tools in all circumstances when genAI tools were used in some way.

To clarify user guidance in this regard, agencies endorsing this mitigation measure could restate the goal/s of AI content flagging and marking and specify the forms of flagging or marking that satisfy the goal/s. According to Wittenberg et.al., for example, AI content labelling could have a process-based goal, i.e., to reveal 'the process by which a given piece of content was created or edited' (whether genAI tools were used or not) or an impact-based goal, i.e., to decrease 'the likelihood that content misleads or deceives' its recipient.[76] Depending on the goal, the specific actions required from users may or may not be deemed meaningful. For example, if the goal is impact-based, then it may not be sufficient to simply have a general disclosure that AI was used. It will be more important to consider who is the recipient who may be misled and under what circumstances and to tailor markings or warnings that will be helpful to such recipients. Even where the goal is process-based, if everyone simply included a general disclosure of AI usage in all documents, then it becomes a meaningless tick-box exercise.

5.2 Addressing Impacts on Team Dynamics and Human Abilities

The introduction and integration of LLM tools into the workplace could have significant impacts on team dynamics and long-term implications on valued human abilities.[77] These risks are what Kaminski may regard as 'unquantifiable' or difficult to quantify and will likely be missed by a risk regulation approach, much less in one focused on end-users. Incorporating these risks into existing risk regulation means research into workers' experiences of genAI's impact on work relations, including equity, and valued human abilities should be incorporated into future assessments of risks of using genAI tools at work. How this will or can be done is unclear.

The likely mitigation measures could involve setting workplace policies. Ideally, the boost or enhancement of worker productivity derived from AI should take place in an equitable manner and not leave anyone, particularly those already socially disadvantaged, in a situation where they may be judged as less productive. Realistically, however, not every employee will have access to or will find LLM assistance helpful while others, particularly those with pre-existing AI skills and knowledge, will be able to benefit disproportionately from such assistance. Individual differences and attitudes towards AI upskilling thus may inevitably have an impact on whether a new digital divide arises or is entrenched within organisations.

While employees' access to proprietary AI remains on a selective basis, it is crucial for there to be a well-justified policy for distributing access to expensive genAI tools, and for such policy to anticipate and address the impacts on team dynamics. Trainings could help if they created genuine opportunities to equalise development of AI skills and knowledge across the workforce. Agencies' efforts towards promoting equity in the workplace could include decisions regarding the use of LLMs to improve the performance

[75] Office of Digital Government, 'Large Language Models: WA Public Sector Guidance' (Government of Western Australia 2024).

[76] Chloe Wittenberg and others, 'Labeling AI-Generated Content: Promises, Perils, and Future Directions' (2024) MIT Exploration of Generative AI. The authors discuss 'AI content labelling' in the context of social media, new sites and search engines where it is also commonly proposed as a mitigation measure against misinformation through genAI.

[77] See also, Hao-Ping (Hank) Lee and others, 'The Impact of Generative AI on Critical Thinking: Self-Reported Reductions in Cognitive Effort and Confidence Effects from a Survey of Knowledge Workers' (ArXiv, 13 January 2025).

of disadvantaged groups within the workplace. There could be priority or emphasis given to supporting disadvantaged groups in the LLM allocation policy, user training, and 'communities of practice'. User training and education could furthermore be used to manage and address concerns about the changing requirements of the workplace in terms of the cognitive abilities that are valuable to have.

5.3 Developing Further Responsibilities and Collaborations

More significantly, many of our respondents were aware that new use cases for genAI in government were a near eventuality and these new uses will pose new challenges to the focus on user responsibility to mitigate genAI. The risks entailed may not be addressed through the options discussed above. Rather, they will require the government to review the distribution of responsibilities among users, deployers and developers and create more collaboration across the genAI value chain.

Chatbots as interfaces between agencies and the public will heighten attention to misinformation, bias and information oversharing compared to the use of Copilot as personal assistant. The public's use for the chatbot, i.e., to obtain information, advice or service from the government, will be very different from public officers' appropriate use for genAI tools for assistance with drafting emails or presentation slides. Thus, user guidance for the public will also be very different. For example, instructing the public to review or verify information from government chatbots would conflict with the assumption that the public is supposed to trust the government. Moreover, it will not be realistic or reasonable to expect the general public to restrain themselves from oversharing information like sensitive personal information with the chatbot. Even when this guidance is provided, the public cannot be expected to exercise the same level of conscientiousness in complying with such guidance compared to public officers who are subject to internal governance mechanisms that ultimately enforce such guidance.

Rather than focusing on users, the government should articulate duties and requirements addressed to deployers and developers. In the case of government chatbots, mitigation will likely include the moderation of LLM's responses before they are sent out to public end-users to prevent or attenuate unsafe or harmful responses. Human moderators or supervisors of the government chatbot will likely be required to ensure there is a human-in-the-loop. Unless moderation mechanisms are adequate to ensure safety, using LLMs to speed or scale up delivery of government information, advice or service will be of doubtful overall benefit to government and the public. However, the adequacy and security of moderation is not settled by the mere fact that it introduces a human-in-the-loop. Rather, the government will contend with questions about the qualification of human moderators, i.e., who should have responsibility to supervise the chatbot. It will also contend with the nature and quality of human-AI system interaction which can imply that the AI system should be designed to facilitate successful human supervision. Thus, developer requirements regarding the transparency and explainability of the AI system's design will have to be developed.[78]

As one of our respondents argue, before we see genAI systems deployed by Australian government agencies as public-facing chatbots, it is likely that agencies will first experiment internally with genAI systems to automate certain internal processes. That is, agencies may try to deploy LLMs to power internal chatbots that serve as interfaces between staff requesting action and internal units responsible for providing staff with advice or decisions. These uses will imply heightened risks associated with increased automation of

[78] Agencies will do well to review issues involving moderation in other contexts. In the context of social media, Australia and various international governments have grappled with the failings of content moderation by social media companies to stem misinformation and other harm to public end-users. See, e.g., Australian Communications and Media Authority, 'Report to Government on the Adequacy of Digital Platforms' Disinformation and News Quality Measures' (2023) https://www.acma.gov.au/report-government-adequacy-digital-platforms-disinformation-and-news-quality-measures accessed 12 August 2024; See, on human and automated methods of content moderation, Anna Veronica Banchik, 'Disappearing Acts: Content Moderation and Emergent Practices to Preserve at-Risk Human Rights–Related Content' (2020) 23 Information, Communication & Society 1339;New Media & Society, UN Special Rapporteur on the Protection and Promotion of the Right to Freedom of Opinion and Expression and others, 'Joint Declaration on Freedom of Expression and Gender Justice'; Alessandra Gomes, Dennys Antonialli and Thiago Dias Oliva, 'Drag Queens and Artificial Intelligence: Should Computers Decide What Is "Toxic" on the Internet?' (*InternetLab*, 28 June 2019) https://internetlab.org.br/en/news/drag-queens-and-artificial-intelligence-should-computers-decide-what-is-toxic-on-the-internet/ accessed 11 July 2023.

decision-making. AI assurance frameworks will challenge risk managers in this instance to address risks through greater fine tuning by deployers or by engaging developers regarding the model's training.

Another potential use of genAI tools is by aiding teams as a more or less *independent team member* or *agent*, for example, as a moderator of meetings or as a project manager.[79] Comparable uses exist or are being proposed in the education sector where genAI agents act as tutors.[80] This would go beyond merely providing individual assistance to public officers in their day-to-day tasks. To be valuable to organisations, the genAI *agent* must be able to perform its tasks with a greater degree of independence (i.e., with less need for human direction implied by constant prompt engineering than current uses) but still do so safely. As with government chatbots, the associated risks will not be adequately addressed through end-user guidance alone. Almost certainly, it requires deployers to engage in significant finetuning of pre-trained models in cooperation with developers.

With regard to high-risk applications of AI, Hacker, Engle and Mauer (2023) have suggested that regulation could be designed on the premise that the entire AI value chain is responsible for harm.[81] This concept implied that actors across the AI value chain should cooperate to avoid or minimise harm and account for harm that occurs. Consequently, they suggested that the public interest in avoiding or accounting for harm in high-risk AI applications should surmount certain private interests of AI actors. Significantly, while respecting the intellectual property interest of developers in keeping their proprietary models under wraps, under certain circumstances or conditions, they argued that it should be possible to allow model inspection by deployers and/or users who have legitimate interests as collaborators in risk management (or in the case of legal accountability, as co-parties with joint and several liability). Hacker, Engle and Mauer (2023) draw inspiration from the evidentiary rules in the EU AI Act as well as certain practices in discovery procedures under US law.

Greater use of auditing could make genAI risk regulation more robust. Mökander et.al. (2023), who focused on governance mechanism for LLMs, have proposed that AI risk managers could benefit from three levels of auditing[82], namely, as applied to the technology provider's governance (governance audit), the pre-trained models themselves before the models are put to use (model audit), and the actual applications of LLMs (application audit). Their suggestion improves drastically on current practice of relying on vendor assurance. Innovatively, they further proposed that these audits be coordinated and structured so that audits on one level 'become inputs for which audits in other levels must account'. They theorise that governance audits could contribute to risk regulation, particularly of cybersecurity risks.[83] As Mökander et.al. (2023) themselves acknowledge, however, none of these theoretical uses and benefits of auditing for risk regulation will be realised unless there is practical compulsion for developers, deployers and users alike to be responsible for potential AI harms and cooperate in their mitigation.[84] The foregoing discussion suggests that the notions of whole-of-AI-value-chain responsibility and collaboration along the AI value chain could enrich Australian

[79] Already, Microsoft is marketing newer versions of Copilot for these purposes. See, Jared Spataro, 'New Agent Capabilities in Microsoft Copilot Unlock Business Value' (*Microsoft 365 Blog*, 21 May 2024) https://www.microsoft.com/en-us/microsoft-365/blog/2024/05/21/new-agent-capabilities-in-microsoft-copilot-unlock-business-value/ accessed 29 July 2024.

[80] Bernard Marr, 'Online Education and Generative AI: Welcome to the Age of Virtual AI Tutors' (*Forbes*,6 June 2024) https://www.forbes.com/sites/bernardmarr/2024/06/06/online-education-and-generative-ai-welcome-to-the-age-of-virtual-ai-tutors/ accessed 20 August 2024.

[81] Hacker, Engel and Mauer (n 25) 1117.

[82] "[A]uditing is a systematic and independent process of obtaining and evaluating evidence regarding an entity's actions or properties and communicating the results of that evaluation to relevant stakeholders. Three ideas underpin the promise of auditing as an AI governance mechanism: that procedural regularity and transparency contribute to good governance; that proactivity in the design of AI systems helps identify risks and prevent harm before it occurs; and, that the independence between the auditor and the auditee contributes to the objectivity and professionalism of the evaluation" (Mokander et.al. 2023, p. 1)

[83] Mökander and others (n 43).

[84] Emanuel Moss and others, 'Assembling Accountability: Algorithmic Impact Assessment for the Public Interest' (Data & Society 2021) similarly argued that impact assessments – or for our purpose, risk assessments – will fail to have the regulatory effect it was designed to have if they imply no accountability on specific actors. In turn, accountability only results from impact assessments when the latter is legitimised "through legislation or within a set of norms that are officially recognised and publicly valued."

risk regulation of genAI. They may well be necessitated by scrupulous adherence to AI assurance frameworks in higher risk applications.

6. Conclusion

Our examination of the Copilot deployment in Australian government agencies revealed that risk regulation as actually practiced in relation to genAI in government in the Australian context needs improvements that gesture beyond a 'light touch' approach.

Agencies have generally relied on end-users to manage or mitigate risks, particularly misinformation and social inequality-related risks. Even when Copilot usage is limited to personal assistance tasks, however, end-users could be challenged by time and productivity pressures and the lack of clarity regarding the guidance on content flagging and marking, and the avoidance of genAI usage in decision-making contexts. It is possible to strengthen end-user-focused mitigation measures currently emphasised in basic genAI end-user guidance documents. However, we have also highlighted genAI's impact on team dynamics and long-term implications on human abilities. These risks are largely ignored in risk regulation arguably because of the difficulty or impossibility of quantifying them. Finally, our case study has shown that imminent applications of genAI by Australian government agencies, particularly, internal or public-facing government chatbots, will require drastically more robust measures than what are available within 'light touch' risk regulation. These measures involve defining and requiring the responsibility, accountability and collaboration of actors in the AI value chain beyond the end-users.

6.1 Disclosure statement

The authors report there are no competing interests to declare.