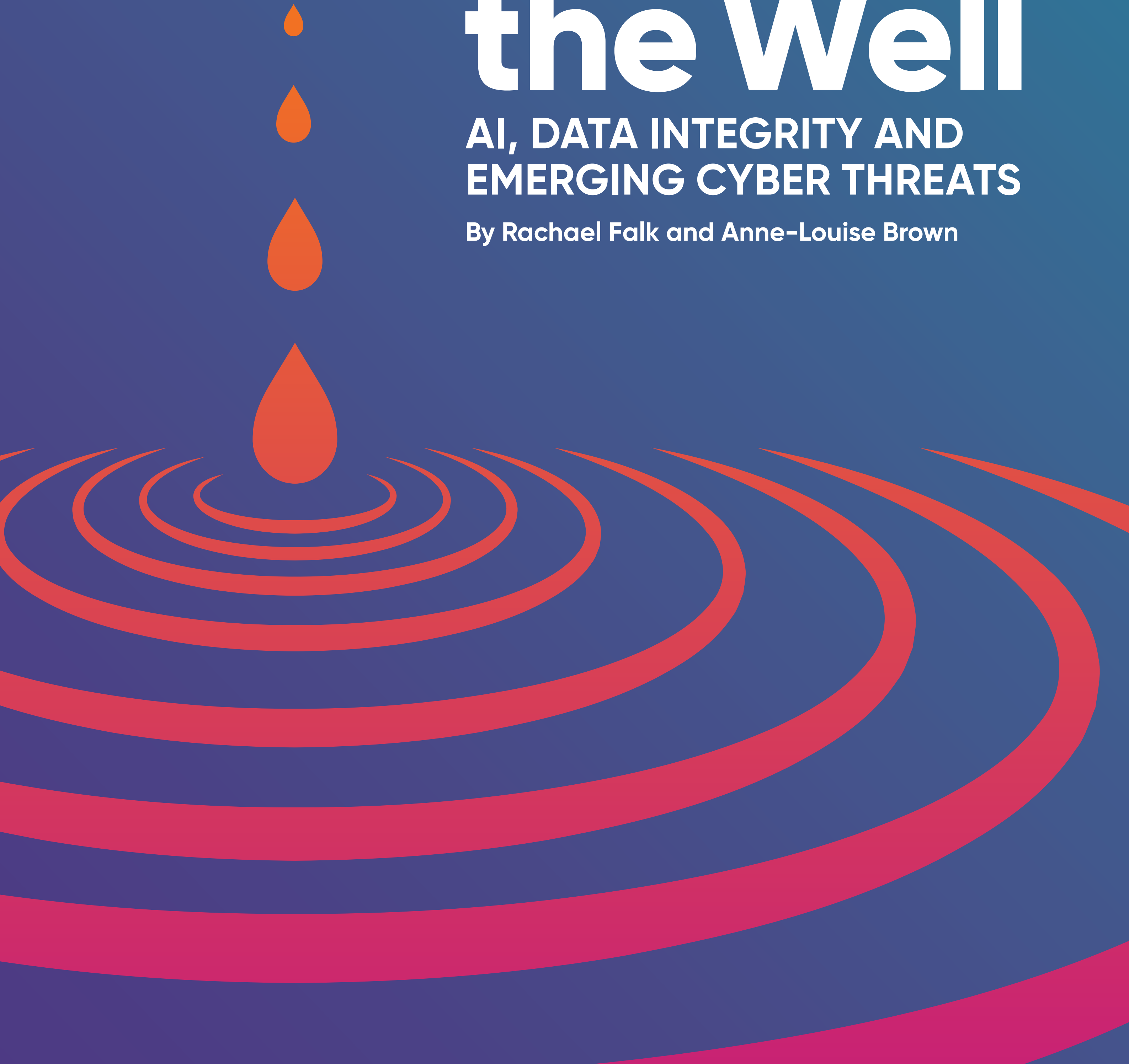
CYBER SECURITY
COOPERATIVE
RESEARCH
CENTRE

cybersecuritycrc.org.au

Poison the Well

AI, DATA INTEGRITY AND EMERGING CYBER THREATS

By Rachael Falk and Anne-Louise Brown



CYBER SECURITY
COOPERATIVE
RESEARCH
CENTRE

WITH THE SUPPORT OF

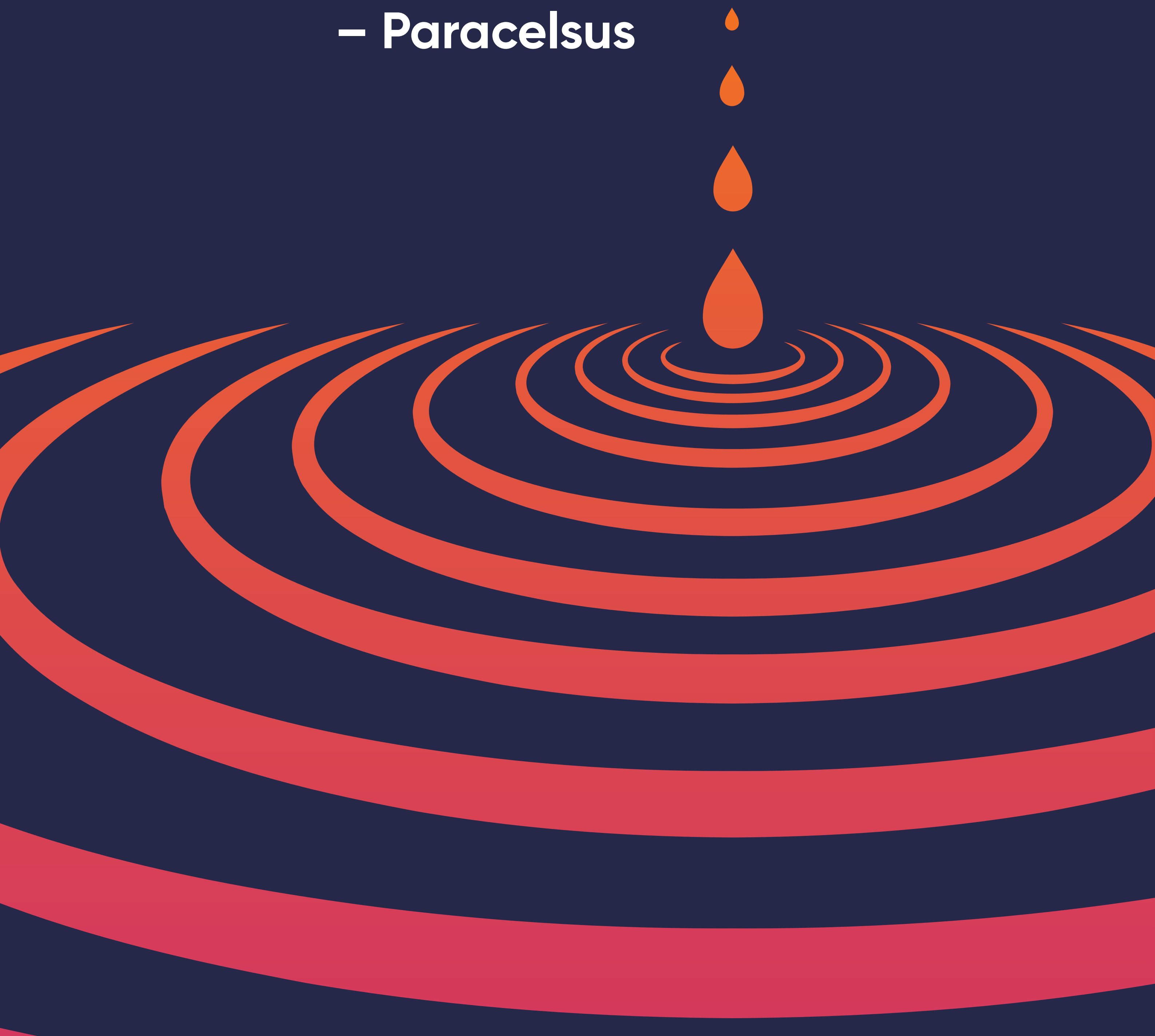
Australian Government
Department of Industry,
Science and Resources

AusIndustry
Cooperative Research
Centres Program

Disclaimer: This publication is designed to provide accurate and authoritative information in relation to the subject matter covered. It is provided with the understanding that the publisher is not engaged in rendering any form of professional or other advice or services. No person should rely on the contents of this publication without first obtaining advice from a qualified professional.

"The dose
makes
the poison."

– Paracelsus



CONTENTS

INTRODUCTION

Artificial intelligence (AI) is the next frontier. It is the printing press of the 21st century, digital penicillin, the lightbulb moment that promises to change our collective futures forever, for better or for worse.

AI has the potential to do much good, creating new efficiencies and solutions that change the way we live and work forever. It has the power to accelerate scientific development and support climate action, disaster prevention and public service delivery.[1] However, the rise of AI also comes with significant risks, including labour market displacement, digital poverty, and threats to privacy and safety.[2]

Another key area of risk is cyber security. This is an area that has been explored in depth in relation to specific threat types, with broad commentary regarding deepfakes, AI-enabled cyber attacks on autonomous vehicles, tailored and sophisticated large-scale phishing exploits and the spread of disinformation.[3] However, less attention has been paid to other less visible, but just as dangerous, AI-cyber threat vectors.

In cyber security, confidentiality, integrity and availability are known as the 'CIA triad' and, for cyber security to be effective, each factor must be managed.[4] And in relation to AI models, which rely on data for training, data integrity is key. It is critical that data sets can be trusted.

This policy paper will explore two key emerging AI-related cyber threat types with the potential to result in serious security, societal and personal harms. They are:

- Data poisoning attacks on AI data training sets; and
- Cyber threats associated with human labelling of AI data sets.

These threat vectors will be explored in the context of Generative AI (GAI), namely large language models (LLMs). Furthermore, this paper seeks to provide an easily understood definition of what AI is, what LLMs are, and how AI is trained.

Finally, four key recommendations are provided to help Australian policy makers prepare for and manage these emerging risks. These are:

- Oversight, transparency and governance measures be introduced for AI training data sets;
- Domestic harmonisation with international AI regulatory regimes;
- Leveraging Australia's Modern Slavery regime for enhanced oversight of AI supply chains; and
- A focus on research investment targeted at establishing Australia as a leader in the identification and mitigation of emerging AI-related cyber threats.

1. Harnessing the power of AI and emerging technologies, OECD
2. Ibid.
3. AI-enabled future crime, Dawes Centre for Future Crime, University College London
4. Executive Summary — NIST SP 1800-26 documentation

WHAT IS AI?

In the simplest of terms, AI refers to the use of machine-driven technologies that can perform tasks previously requiring human intelligence.

However, as highlighted by Kay Firth-Butterfield, the Head of AI and Machine Learning at the World Economic Forum (WEF), current forms of AI are not 'intelligence' per se, but prediction. She observes that: "For the most part, (AIs) can still only do one task very well at a time. This is not commonsense and is not equivalent to human levels of thinking that can facilitate multi-tasking with ease. Humans can take information from one source and use it in many different ways. In other words, our intelligence is transferable—the 'intelligence' of machines is not".[5]

Furthermore, due to the diversity and complexity of AI systems, the formulation of a universal overarching definition of 'AI' has been difficult to articulate.[6] Currently, on the international stage, the OECD's definition of AI has been most broadly adopted and, for the purposes of this paper, this definition of 'AI system' has been applied.

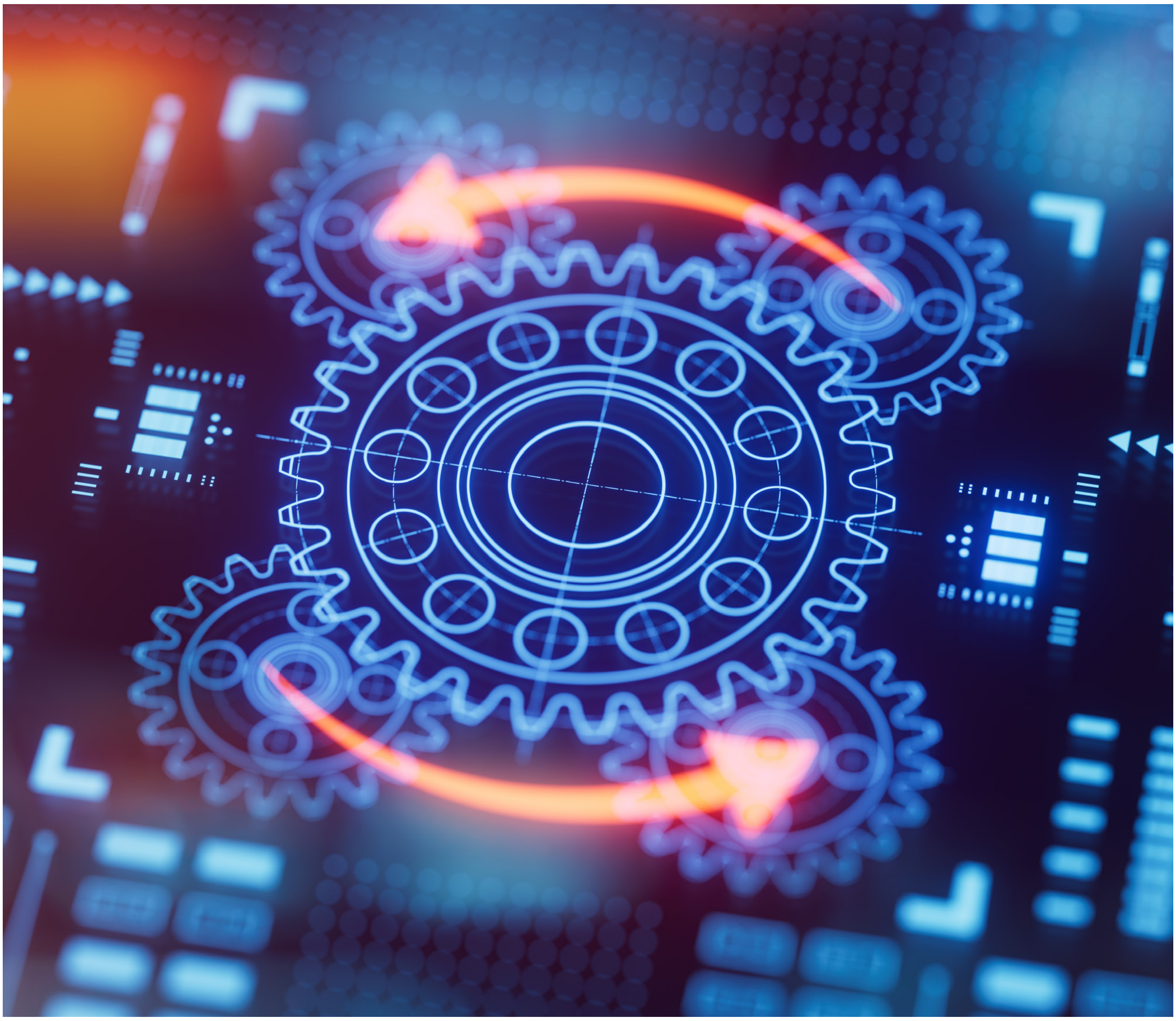
ACCORDING TO THE OECD:

"An AI system is a machine-based system that is capable of influencing the environment by producing an output (predictions, recommendations or decisions) for a given set of objectives. It uses machine and/or human-based data and inputs to (i) perceive real and/or virtual environments; (ii) abstract these perceptions into models through analysis in an automated manner (e.g., with machine learning), or manually; and (iii) use model inference to formulate options for outcomes. AI systems are designed to operate with varying levels of autonomy".[7]

The key enabling technology underpinning AI systems is machine learning (ML). As explained by Monash University's Dr Campbell Wilson in his essay, *Living with AI*, a ML system 'teaches' a computer how to solve specific problems through analysis of a defined data set.[8] Furthermore, supervised ML steers the machine towards the responses it is designed to produce, training it to respond correctly. Once trained on a defined data set, new data to which an AI system has not been exposed to can be input, with the machine applying its previous learning to this unfamiliar data with the purpose of producing correct outputs.[9]

Recently, GAI models have garnered significant attention, driven by the startling launch of ChatGPT. GAI refers to systems that, unlike traditional AI systems designed to recognise patterns and make predictions, create new content. As previously noted, for the purposes of this paper, focus will be placed on data integrity in LLMs, which are a form of GAI.

[5] What is artificial intelligence—and what is it not? | World Economic Forum
[6] Explainer: What is a foundation model? | Ada Lovelace Institute
[7] AI-Principles Overview - OECD.AI
[8] Wilson, C., Living with AI, Monash University Publishing, 2023 PP 10-11
[9] Ibid.
[10] Safe and responsible AI in Australia



Put simply, LLMs, like ChatGPT, specialise in the generation of human-like text, known as natural language processing.[10] Sejnowski describes them as "pretrained foundational models that are self-supervised and can be adapted with fine-tuning to a wide range of natural language tasks, each of which previously would have required a separate network model".[11] LLMs are trained using massive data sets with billions of parameters, enabling them to undertake tasks like text and code generation and translation on command via, for example, instruction-following chatbots.[12]

More complex than LLMs, multimodal foundation models (MfMs) can process and output multiple data types, including text, images and audio.[13] They use a wider range of data inputs including images, speech, numbers and code, and are trained to identify the relationship between various inputs.[14] As highlighted by the Ada Lovelace Institute, a defining characteristic of MfMs is the vast scale of data and computational resources involved in building them – "they require datasets featuring billions of words or hundreds of millions of images scraped from the internet".[15] Significant concerns have been raised about the propensity for MfMs in the dissemination of large-scale disinformation attacks via deepfake imagery and videos.[16]

[11] Large Language Models and the Reverse Turing Test | Neural Computation | MIT Press
[12] A Gentle Introduction to Open Source Large Language Models | Towards Data Science
[13] Safe and responsible AI in Australia (storage.googleapis.com)
[14] Rapid Response Information Report: Generative AI (atse.org.au)
[15] What is a foundation model? | Ada Lovelace Institute
[16] Deepfakes generation and detection: state-of-the-art, open challenges, countermeasures, and way forward | NSF PAGES

TRAINING AI – THE DATA DILEMMA

For AI to be effective and operate as desired, it requires training. As previously mentioned, this occurs via the use of training data sets.

LLMs are generally 'pre-trained' in an unsupervised manner using huge data sets, learning to identify patterns and structures of language, images, speech, numbers and code. During this phase, data can be edited, added or removed to increase accuracy and reduce bias.[17] Following this process, models are fine-tuned to perform specific tasks, for which a smaller, more specialised data set is employed, with responses to specific tasks resulting from input prompts.[18] This process relies on human-in-the-loop judgement, supervised learning and reinforcement learning, which enables an AI system to learn through trial and error using positive and negative feedback from its actions.[19]

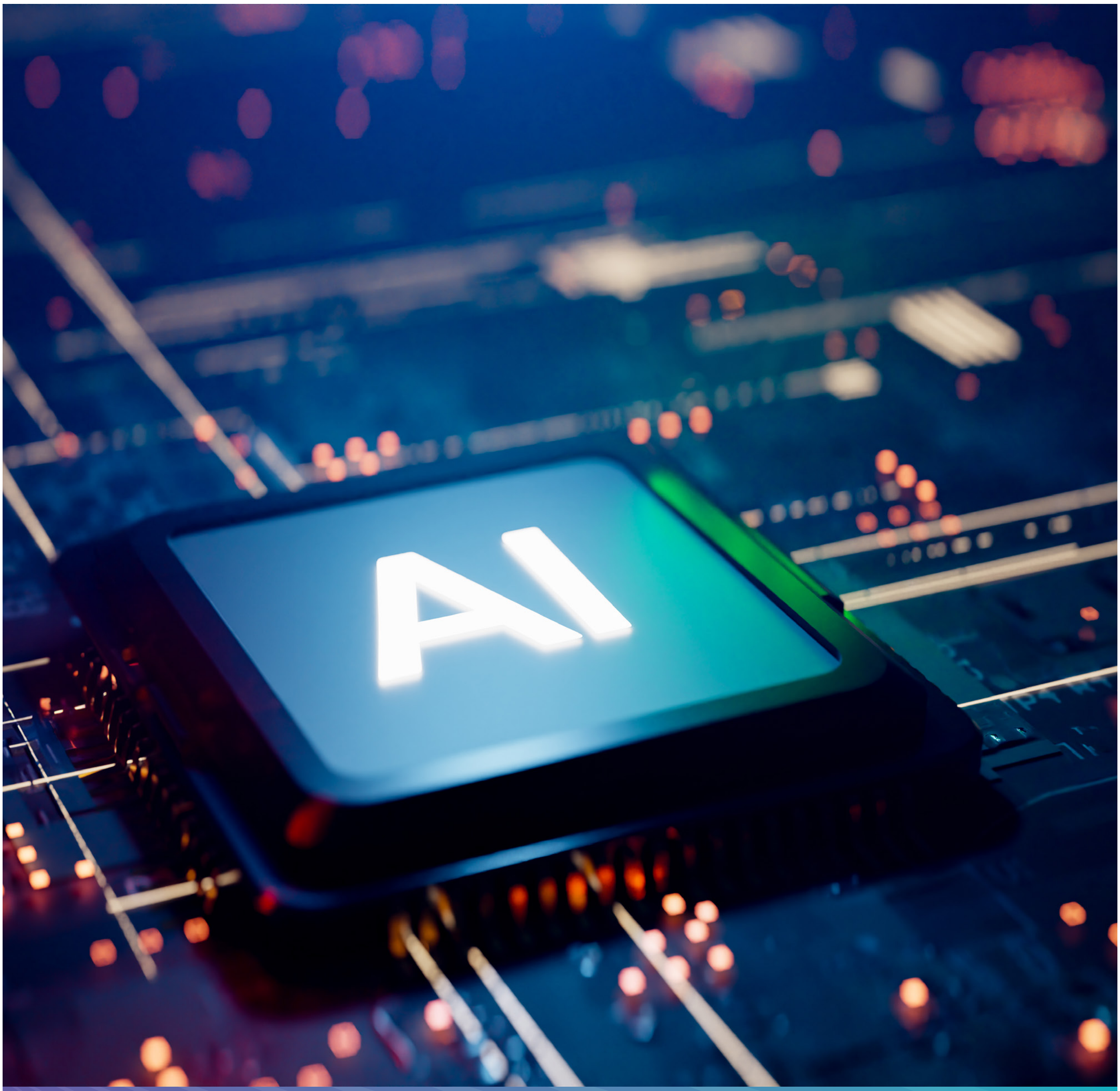
However, as recently highlighted by the UK's National Cyber Security Centre (NCSC), AI models are only as good as the data they are trained on. And this is where things get blurry, because there is often a lack of transparency as to where training data comes from and, therefore, its integrity is a key issue for consideration.

Currently, vast amounts of training data are scraped from the open internet without moderation, which means data sets inevitably include offensive, inaccurate or controversial content,[20] and many LLMs are trained using primary text sources like Wikipedia and news articles.[21] Such data may be incorrect and contain biases like, for example, cultural or political biases. Wilson states that "given the internet is hardy devoid of what could politely be called misinformation, any AI model trained on large swathes of it without careful curation, and training of the model for safety, may learn misinformation and consequently cannot be relied on".[22]

Further, as noted by The Alan Turing Institute's Dr David Leslie: "Responsible data acquisition, handling, and management is a necessary component of algorithmic fairness. If the results of your AI project are generated by biased, compromised or skewed datasets, affected stakeholders will not adequately be protected from discriminatory harm".[23] Ultimately, this means that data sets should be representative to prevent bias, fit-for-purpose in regard to the volume of data required, up to date to ensure currency and relevance, and be appropriate for the desired application.[24] However, current data gathering techniques cannot provide such assurances.

While data cleaning and sanitisation methods are used to help prevent data set inaccuracies and biases, they are not a silver bullet. That is because removing inaccuracies and biases is not purely a technical issue, but also a social one. As noted by the National Institute for Standards and Technology (NIST), "when human, systemic and computational biases combine, they can form a pernicious mixture – especially when explicit guidance is lacking for addressing the risks associated with using AI systems".[25] To help counter these issues, NIST has advocated for a 'socio-technical' systems approach, which would assist in evaluating dynamic systems of bias, their interactions with each other and how they could be reduced.[26]

[17] Rapid Response Information Report: Generative AI (atse.org.au)
[18] Understanding of Large Language Models in detail (used in ChatGPT) | Medium
[19] What is reinforcement learning? - University of York
[20] Thinking about the security of AI systems - NCSC.GOV.UK
[21] We could run out of data to train AI language programs | MIT Technology Review
[22] Wilson, C., Living with AI, Monash University Publishing, 2023 P54
[23] Understanding AI ethics and safety (turing.ac.uk)
[24] Ibid.
[25] There's More to AI Bias Than Biased Data, NIST Report Highlights | NIST
[26] Towards a Standard for Identifying and Managing Bias in Artificial Intelligence (nist.gov)



WHAT IS 'TRUSTWORTHY AI'?

As defined by the OECD, 'trustworthy AI' refers to AI systems that respect human rights and privacy; are fair, transparent, explainable, robust, secure and safe; and the actors involved in their development and use remain accountable.

Such systems consider five key considerations:

1. Inclusive growth, sustainable development and wellbeing
2. Human-centred values and fairness
3. Transparency and explainability
4. Robustness, security and safety
5. Accountability: AI actors should respect the principles and should be accountable for the proper operation of AI systems.

Source: Tools for trustworthy AI: A framework to compare implementation tools for trustworthy AI systems, OECD

WHY WOULD MALICIOUS ACTORS TARGET DATA?

In the world of AI, data accuracy and integrity are everything.

If data is incorrect or is biased it means an AI system will not be fair or accurate, ultimately making it untrustworthy. Therefore, any threat to AI data inputs presents a threat to the integrity of an AI system itself and, more worryingly, could have serious societal impacts. Hence, cyber attacks aimed at manipulating AI data sets must be considered as a serious emerging cyber threat vector of which policy makers, AI developers and organisations need to be increasingly aware.

Such an approach is especially important in relation LLMs, which are being adopted rapidly by citizens, organisations and governments. If the data used to train such systems used by governments and large private sector organisations was compromised, this could have serious ramifications for citizens and consumers, from false and misleading results being disseminated right through to serious output bias and potentially the spread of disinformation. Furthermore, at a time when AI systems offer significant savings and efficiencies, undermining of trust in these systems could stall their development and hinder innovation.

This paper explores two of potential threat vectors that could threaten the integrity of AI datasets – data poisoning and human threats associated with labelling of GAI data.

Data poisoning

Data poisoning attacks occur when an AI system is being trained. Such attacks involve the input of malicious, biased or incorrect data into an AI data training set, resulting in the model learning false patterns and making incorrect connections. Hence, when a poisoned model is deployed, it will produce incorrect outputs that could enable an attacker to bias decision-making towards a particular outcome, which could result in real-life harms.[27]

Data poisoning is a form of adversarial machine learning (AML) that is "very powerful and can cause either an availability violation or an integrity violation".[28] Accordingly, NIST researchers have identified four key data poisoning types that apply to AI systems: availability poisoning; targeted poisoning; backdoor poisoning; and model poisoning.[29]

In relation to data poisoning and the impetus to prevent it, there are also commercial concerns for developers. Research indicates that securing LLMs against data poisoning is impossible "without significant sacrifices in performance due to their inevitable memorisation and need for memorisation for good performance".[30] In short, preventing data poisoning of LLMs has the potential to be expensive and reduce system performance, potentially impacting commercial viability.

Data poisoning attack types

AVAILABILITY POISONING:

The entire ML model is corrupted in an availability attack, resulting in model misclassification on the majority of testing samples and a significant reduction in model accuracy, rendering it unusable.[31]

TARGETED POISONING:

An attack is localised to one or a small number of testing samples, making it difficult to detect.[32]

BACKDOOR POISONING:

An attacker introduces backdoors into a set of training examples, which trigger the model to misclassify samples with the same backdoor pattern during testing.[33]

MODEL POISONING:

These attacks attempt to directly modify the trained ML model to inject malicious code into the model.[34]

[27] Adversarial Attacks On AI Systems (forbes.com)
[28] Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations (nist.gov), P20
[29] Ibid., PP 20-26
[30] LLM censorship: A machine learning challenge or a computer security problem?, P8

DATA SCRAPING AND BLIND TRUST – A DANGEROUS COMBINATION

Data poisoning attacks are not the stuff of fiction – they can be cheap and easy to execute, as illustrated by a team of researchers from Google, Nvidia, Robust Intelligence and ETH Zurich.[35]

In their paper, *Poisoning Web-Scale Training Datasets is Practical*, the researchers found that for just USD$60 they could purchase domains and fill them with images of their choice, which were scraped into large data sets. Furthermore, they were able to edit and add sentences to Wikipedia entries, poisoning data that formed part of an AI model's data set.[36]

As stated by the researchers: "We introduce two novel poisoning attacks that guarantee malicious examples will appear in web-scale datasets used for training the largest machine learning models in production today. Our attacks exploit critical weaknesses in the current trust assumptions of web-scale datasets: due to a combination of monetary, privacy, and legal restrictions, many existing datasets are not published as static, standalone artifacts. Instead, datasets either consist of an index of web content that individual clients must crawl; or a periodic snapshot of web content that clients download. This allows an attacker to know with certainty what web content to poison (and, as we will show, even when to poison this content)".[37]

Two methods of data poisoning were used for the research: split-view data poisoning and front-running data poisoning. Split-view data poisoning operates upon the understanding that while the index of a training dataset cannot be altered, content of URLs in the dataset can, enabling "an adversary who can exert sustained control over a web resource indexed by the dataset to poison the resulting collected dataset collected by the end-user".[38] Front-running data poisoning occurs when an adversary is able to alter web content in a very short span of time during which modifications cannot be detected. This is possible if a malicious actor is able to accurately predict when such web content will be accessed for a dataset snapshot. Using the example of Wikipedia, which is used extensively for AI data sets, the researchers found they could predict the timing of a data snapshot being captured "down to the minute".[39] This allowed them to insert inaccurate data in the minutes before a data snapshot was taken, during which there was not enough time for the inaccurate data to be amended by Wikipedia. As a result, incorrect data was captured and fed into AI training data sets.

Ultimately, the researchers concluded LLMs are vulnerable to attack before deployment, meaning "ML researchers must reassess the trust assumptions they place in web-scale data and begin exploring solutions that do not assume a single root of trust".[40]

[31] Poisoning Attacks Against Machine Learning: Can Machine Learning Be Trustworthy? (ieeecomputer.org)
[32] Ibid.
[33] Ibid.
[34] Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations (nist.gov), P26
[35] Poisoning Web-Scale Training Datasets is Practical
[36] Three ways AI chatbots are a security disaster | MIT Technology Review
[37] Poisoning Web-Scale Training Datasets is Practical
[38] Poisoning Web-Scale Training Datasets is Practical
[39] Ibid.
[40] Ibid.

HUMAN AI LABELLING

While AI development continues unabated, a significant humanitarian and cyber security issue is emerging in some developing nations, where 'AI sweatshops' are emerging to fulfil the voracious appetite for GAI data labelling.[41] [42] [43]

Kenya, Uganda, Philippines and Venezuela are among the developing nations to which AI labelling is being outsourced by large tech companies. A recent Time article highlighted the low pay and tough conditions Kenyan LLM labelling workers have endured, where OpenAI, creators of ChatGPT, through a third-party provider, pay workers about USD$2 an hour to label confronting content.[44] These workers have been tasked with labelling offensive content so OpenAI can develop an AI system to detect and remove such content from ChatGPT. The article states: "To get those labels, OpenAI sent tens of thousands of snippets of text to an outsourcing firm in Kenya, beginning in November 2021. Much of that text appeared to have been pulled from the darkest recesses of the internet. Some of it described situations in graphic detail like child sexual abuse, bestiality, murder, suicide, torture, self-harm, and incest".[45]

In the Philippines, it is estimated about two million people perform AI-labelling 'crowdwork', with workers raising concerns about low-wage exploitation.[46] And similarly, in Venezuela, many workers turned to the gig-economy to label GAI data as a source of income as the nation's economy crumbled, creating an AI crowdwork hub.[47]

Understandably, ethical dilemmas have been raised in relation to potential worker exploitation. To date, however, the possible cyber security implications of using such a workforce has not been considered.

This paper raises the spectre that poor workers and corrupt officials in developing nations may be particularly vulnerable to coercion by malicious parties, who could employ financial incentives to use such a workforce to manipulate the labelling of LLM training data. Such attacks, known as label poisoning and input attacks, occur when an adversary injects mislabelled or malicious data into an AI training set to influence its behaviour and alter its outputs. In effect this could see offensive materials or text sequences, like that related to child abuse material or rape, intentionally incorrectly labelled as something harmless, like a tree, cat or bag. If this was to occur at scale across a range of different damaging scenarios– and research indicates only 0.01% of training data needs to be poisoned to be effective[48] – the impacts on LLMs could be serious and the implications for society damaging. Most importantly, such an attack would have a deleterious impact on perceptions of GAI at a social and cultural level, impacting the positive economic and societal effects these technologies can affect.

While there is no evidence that such attacks have yet occurred, other forms of cyber-based poisoning attacks, like malware, are well known threat vectors. Therefore, with potential attacks on AI training data sets seemingly inevitable, it is an emerging issue that policy makers must consider as AI technologies, notably GAIs, become more ubiquitous.

> ***"Understandably, ethical dilemmas have been raised in relation to potential worker exploitation. To date, however, the possible cyber security implications of using such a workforce has not been considered. "***

[41] Scale AI's Remotasks workers in the Philippines cry foul over low pay - The Washington Post
[42] How the AI industry profits from catastrophe | MIT Technology Review
[43] OpenAI Used Kenyan Workers on Less Than $2 Per Hour: Exclusive | Time
[44] Ibid.
[45] Ibid.
[46] Scale AI's Remotasks workers in the Philippines cry foul over low pay - The Washington Post
[47] How the AI industry profits from catastrophe | MIT Technology Review
[48] Poisoning Web-Scale Training Datasets is Practical

WHAT ABOUT FOREIGN INTERFERENCE?

In his 2023 Threat Assessment, Australia's Director-General of Security Mike Burgess stressed that in an increasingly complex global environment, foreign interference was the Australian Security and Intelligence Organisation's most pressing concern.[49] He stated that: "Threats are increasingly intersecting, emerging from new places and blurring traditional distinctions. A foreign power can simultaneously be interfering, spying, and setting up for sabotage".[50] Through such a prism, AI systems must be considered a potential tool for foreign interference now and into the future.

Such sentiment has also been expressed by RAND Organisation, with a recent paper highlighting the potential national security threat posed by GAI. The paper states it is an issue that needs to be tackled urgently "in a global environment where foreign interference by totalitarian states is an ever-present issue".[51]

While there is no evidence that GAI labelling manipulation is occurring in the countries mentioned in this paper, there is clear proof that several authoritarian states, notably China and Russia, have significant interests and presence in African and South American nations.[52] [53] [54] Furthermore, there is also a myriad of evidence to indicate endemic public service corruption in Kenya, Uganda, Philippines and Venezuela. [52] [53] [54]

In relation to GAI labelling operations, such factors give weight to the hypothesis that by-proxy these nations could be used for foreign interference operations focussed on AI data manipulation via human labelling. Such interference could result in significant biases and misinformation being injected into GAI data sets. More worryingly, such a tactic could be applied via pressure on a corrupt public service, not through direct workforce coercion. At a global level, this is an issue that needs to be considered urgently as discussions regarding AI regulation and global norms for AI systems continue.

[49] Director-General's Annual Threat Assessment | ASIO

[50] Ibid.

[51] The Rise of Generative AI and the Coming Era of Social Media Manipulation 3.0: Next-Generation Chinese Astroturfing and Coping with Ubiquitous AI (rand.org)

[52] Russia overtakes China as leading arms seller in sub-Saharan Africa

[53] China-Venezuela Economic Relations: Hedging Venezuelan Bets with Chinese Characteristics | Wilson Center

[54] What Does the Ukraine-Russia War Mean for Kenya? – KIPPRA

[55] CPI 2021 for Sub-Saharan Africa: Amid democratic... - Transparency.org

[56] A look at how corruption works in the Philippines | Inquirer Business

[57] Venezuela - Transparency.org

CONCLUSION AND RECOMMENDATIONS

There are many gains to be made through the widespread adoption of AI systems in society – but there are also many risks. Cyber security is of vital importance globally and, in considering the rise of AI, must be a key consideration.

This paper highlights two potential emergent forms of cyber attack that could impact AI systems and, therefore, erode public trust in these technologies more broadly. That means, while there is still time, steps must be taken to ensure the integrity of AI training data sets.

Therefore, this paper makes four key recommendations for Australian policy makers to consider as discussion regarding domestic regulation of AI systems continues.

Oversight, transparency and governance measures be introduced for AI training data sets

As highlighted, the data collection processes for AI data sets are currently opaque, with a lack of oversight in place to govern how data is collected and how it is used. As noted by NIST researchers, most of the companies developing LLMs do not release detailed information about the data sets that have been used to build their models.[58] And in its *Safe and responsible AI in Australia – Discussion paper*, the Federal Government noted the problems caused by inaccurate data.[59] Therefore, the introduction of oversight, transparency and governance mechanisms for AI training data sets should be considered.

To enhance transparency and accountability processes for AI systems, researchers from the Ada Lovelace Institute have suggested developers make available model cards and datasheets, which provide information on the data a system was trained on. It is suggested that these “have the potential to increase transparency and accountability within the machine learning community, mitigate unwanted societal biases in machine learning models, facilitate greater reproducibility of machine learning results, and help researchers and practitioners to select more appropriate datasets for their chosen tasks”.[60]

While achieving oversight of AI training data sets is difficult - and fraught - it has been addressed in Article 10 of the forthcoming European Union AI Act (the Act). Article 10 states that training, validation and testing data sets for “high risk” AI systems shall be subject to appropriate data governance and management practices, be relevant, representative, free of errors and complete, and take into account, to the extent required by the intended purpose, the characteristics or elements that are particular to the specific geographical, behavioural or functional setting.[61] While such an approach is broad and likely to be amended prior to the Act coming into force, Article 10 represents a pragmatic first step in achieving regulatory oversight.

In the US, a number of AI companies have made voluntary commitments to enhance the safety, security and transparent development of AI technologies. These commitments include the undertaking of internal and external security testing of AI systems before their release and independent vulnerability assessments. In the absence of regulation, such an approach could be applied in Australia while steps towards stronger safeguards are developed.

[58] Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations (nist.gov)
[59] Safe and responsible AI in Australia
[60] Expert explainer: Allocating accountability in AI supply chains | Ada Lovelace Institute
[61] EUR-Lex - 52021PC0206 - EN - EUR-Lex (europa.eu)

Harmonisation with international regimes is essential

In Australia, discussions regarding AI regulation remain in their infancy. While the Federal Government's *Safe and responsible AI in Australia – Discussion paper* has raised the spectre of future regulation, this will take time. Therefore, Australia is in a good position to learn from regulatory activity in other countries and ensure that moves towards domestic regulation are undertaken with a view to international harmonisation and interoperability. This is especially important in relation to standards, which are vital for establishing a stable global AI ecosystem and effective policy enactment.

Therefore, the Australian Government should consider the forthcoming EU AI Act as a test case for how Australia could effectively regulate AI systems. Principles-based approaches, such as those adopted in the US and UK, should also be considered. In particular, the UK's push towards a "contextual, sector-based regulatory framework", through which existing regulators would oversee a set of "central functions" could operate well in an Australian context.[62] Such an approach could be implemented initially through leveraging the *Security of Critical Infrastructure Act (2018)*, before more broad application across the economy.

Leveraging Australia's Modern Slavery regime for enhanced oversight of AI supply chains

Under Australia's *Modern Slavery Act 2018* (the Act), large businesses and other entities operating in the Australian market with annual consolidated revenue of at least A$100 million are required to fulfil the Act's Modern Slavery Reporting Requirement and complete annual Modern Slavery Statements.[63] While the Act is designed to capture serious exploitation, not substandard working conditions or worker underpayment, it does require that captured entities identify and address modern slavery risks, and maintain responsible and transparent supply chains.[64]

With many Australian businesses implementing AI systems into their operations,[65] including GAI, there is an opportunity to monitor AI supply chains via the Act. For example, captured entities that are using AI systems in their operations may be required to provide details of third-parties through which AI technologies are procured and produced. And for businesses developing their own AI systems, details of training data sets and their origin could be provided.

As this paper highlights, there are ethical concerns associated with the outsourcing of human AI labelling to workers in developing nations. Modern slavery in tech supply chains has been a significant global issue, notably in relation to the use of forced Uyghur labour in China,[66] and improved oversight and insight into AI workforces would help businesses avoid modern slavery concerns.

Investment in research to find new ways to mitigate emerging AI-related cyber threats is required

As AI systems, especially GAI, continue to evolve at lightning speed, new ways of dealing with the novel threats they will bring must be considered. As this paper highlights, technical solutions to enhance the accuracy and integrity of massive data sets is a wicked problem that needs to be solved. Therefore, funding specific research focussed on finding a solution to this emerging global issue is essential. Such an undertaking cannot be siloed and there is significant scope for Australia and its allies to work together to find technical solutions to enhance dataset integrity. For example, such collaborative efforts could be accomplished under Pillar Two of the AUKUS agreement.

And while technical research is vital, there is also an acute need to consider the human-in-the-loop. As illustrated by this paper, humans play a key role in helping ensure the integrity of AI systems, but they can also pose a potential cyber security risk. Therefore, research that considers the human factors of AI systems and aims to find solutions for the human outsourcing of unethical and risky data labelling practices should be a priority.

[62] Regulating AI in the UK | Ada Lovelace Institute
[63] Modern slavery | Attorney-General's Department (ag.gov.au)
[64] Ibid.
[65] Australia's AI ecosystem momentum report - CSIRO
[66] Uyghurs for sale | Australian Strategic Policy Institute | ASPI

CYBER SECURITY
COOPERATIVE
RESEARCH
CENTRE