

The Rapid Rise of Generative AI

Assessing risks to safety and security

Ardi Janjeva, Alexander Harris, Sarah Mercer, Alexander Kasprzyk and Anna Gausen

December 2023



About CETaS.....	2
Acknowledgements	2
Executive Summary.....	3
Recommendations	8
1. An Introduction to Generative AI.....	11
1.1 A short history of AI.....	11
1.2 The pace of change	13
1.3 Methodology.....	17
2. Evaluating Political, Digital and Physical Security Risks.....	20
2.1 Political security	21
2.2 Digital security	28
2.3 Physical security	33
2.4 Weighing malicious and incidental sources of risk.....	36
3. Generative AI and Future Intelligence Capabilities: Opportunities and Limitations.....	43
3.1 Enlarging the investigative toolbox: analysis and summarisation.....	43
3.2 Open-source and commercial models: innovation vs risk management.....	52
4. Governance, Policy and Regulation	55
4.1 Signalling and reporting	56
4.2 Prohibition and 'red lines'	66
4.3 Strengths and weaknesses in the legislative environment.....	67
4.4 Global governance	68
4.5 Training, guidance and safeguards.....	74
Case Studies.....	77
Case study 1: OSINT Summarisation	77
Case study 2: Synthetic Data Generation	80
About the Authors.....	84

The image used on the cover and back cover was generated by OpenAI's DALL-E 2.

About CETaS

The Centre for Emerging Technology and Security (CETaS) is a research centre based at The Alan Turing Institute, the UK's national institute for data science and artificial intelligence. The Centre's mission is to inform UK security policy through evidence-based, interdisciplinary research on emerging technology issues. Connect with CETaS at cetas.turing.ac.uk.

This research was supported by The Alan Turing Institute's Defence and Security Programme. All views expressed in this report are those of the authors and do not necessarily represent the views of The Alan Turing Institute or any other organisation.

Acknowledgements

The authors wish to thank all those who took part in a research interview or focus group for this project; they are especially grateful to Alena Frankel, Bertie Vidgen, David C, Robert C, and Sam Stockwell for their valuable feedback on earlier versions of this report.

Executive Summary

This CETaS Research Report examines the implications of generative AI for national security. The findings and recommendations are based on a combination of openly available literature and research interviews with more than **50 experts across government, academia, industry, and civil society**. To our knowledge, this represents the most comprehensive publicly available UK-based study on the national security implications of generative AI.

Generative AI is a form of AI that can generate content such as images, audio, and text based on user suggestions. The multitude of possible generative AI use cases is seen by some as an opportunity to revolutionise the way that individuals interact, and businesses operate. However, from a national security perspective, the forms in which generative AI augments human productivity represent a significant challenge and typify the way that **technology is continually stretching the boundaries of national security**.

The security risks posed by generative AI may be understood as either **augmenting pre-existing societal risks or as posing completely novel risks**. In most cases, generative AI lends itself to the former: security risks like disinformation, fraud, and child sexual abuse material are not novel creations of generative AI but are amplified in speed and scale by the technology such that they may harm a larger proportion of the population than before. Understanding the national security picture in this way should dampen unwarranted hysteria regarding the ‘unprecedented’ nature of the threats posed, while enabling a more targeted focus on the threat areas where generative AI catalyses risk.

Generative AI also offers potential opportunities for use within the national security community. Currently, generative AI tools are **too unreliable and error-prone to be trusted in the highest stakes contexts within national security**. This means they are not ready for use where they are required to make a *decision* or where explainability is required to satisfy accountability and oversight requirements. For those who may want to use generative AI to undermine UK national security, inaccuracy is less important – if a large language model (LLM) underperforms in its generation of deepfakes or in writing malware, the cost of failure to the attacker remains low. But from a defensive perspective, similar errors could lead to significant security breaches. Users’ propensity to overly trust LLMs might lead to a **reluctance to challenge AI-generated outputs**.

The national security and technology discourse has historically focused on understanding threats from *adversaries*; groups or individuals who set out to inflict harm. However, the proliferation of advanced technology to a much wider constituency calls for a **shift in mindset** to account for all the **unintentional or incidental ways in which generative AI can pose national security risks**. This is seen in the range of possible instances of ‘improper AI

adoption' – defined as the inappropriate and misguided attainment and deployment of AI systems. In contexts including critical national infrastructure (CNI), public services, the private sector, and individual 'DIY' experimentation, the fear of missing out on the crest of the generative AI wave may cloud judgments about higher risk use cases.

For explicitly *malicious* generative AI use cases, threats can be understood as falling into one of the three categories of digital, physical, and political security.

Digital security	Physical security	Political security
<u>Cybersecurity</u> By reducing the degree of specialist knowledge required, generative AI can assist the less technically able user in experimenting with novel cyberattack techniques and increasing their sophistication iteratively . Less certain is whether generative AI will enable wholly new types of cyberattack that even the best attackers would not previously have been aware of – the most significant longer-term concern from a national security perspective.	<u>Radicalisation and terrorism</u> The personalised relationships that individuals can now form with AI chatbots , in part due to their constant availability and limitless patience, could alter the radicalisation blueprint. However, there remains a distinctly human element to this process which the current generation of generative AI will be unlikely to replicate soon. Specificity about which stages of the terrorist enterprise generative AI is likely to augment is important – for some groups generative AI may be more useful for <i>glorification</i> than <i>radicalisation</i> .	<u>Political disinformation and electoral interference</u> Generative AI could be a force multiplier for political disinformation. The cumulative effect of generative text, image, video, and audio will exceed the impact that any one of those modalities can have individually. Scale could be significantly enhanced by improvements in usability, reliability, and cost-effectiveness of LLMs, while personalisation could reach new levels of convincingness with more impressive storytelling and individually tailored campaigns. In the hours or days preceding an election, it would be challenging to identify and discredit a malicious AI-enabled information operation.
<u>Targeting and fraud</u> Fraudsters stand to benefit significantly from generative AI. Qualitatively, generative AI can assist fraudsters with more professional-looking, highly targeted spear phishing,	<u>Weapon instruction</u> The generation of publicly accessible but hard-to-find information decreases the degrees of separation from information pivotal to developing and executing an	<u>Surveillance, monitoring, and geopolitical fragmentation</u> Generative AI could play an important role in furthering the global proliferation of technology which adheres to authoritarian standards and

<p>increasing the burden of resilience on potential victims. Quantitatively, developments in autonomous agent frameworks could enable wide scale automation of fraud attempts. Improvements in the domain of voice cloning are an area of particular concern in the fraud context.</p>	<p>attack plan. This risk is exacerbated if web APIs permit the connection of large pretrained models into physical systems, which are allowed to take direct actions in the world. Nonetheless, in the biochemical weapons context, there is a significant technical leap from prompting a chatbot to synthesising lethal materials, which limits the utility of generative AI for low-skill actors.</p>	<p>values, aiding attempts to enforce single versions of historical truth for future generations. Democracies may be more vulnerable to the exploitation of the creative characteristics of generative AI systems than autocracies. This emphasises the need to understand the cultural and behavioural aspects to generative AI use around the world.</p>
<p><u>Child sexual abuse material</u></p> <p>The proliferation of AI-generated CSAM is a significant concern for law enforcement agencies. The difficulty of distinguishing 'real' from 'fake' images will continue to increase and pose the challenge of false negatives slipping through the net. At the same time, there is a false positive risk where law enforcement investigates images created of children who have not been physically abused, diverting scarce resources away from those who have.</p>		

Despite their unreliability in very high stakes national security contexts, generative AI does offer various opportunities for national security and law enforcement agencies. In the intelligence analysis context, the role of generative AI is best understood as **enhancing individual productivity**. Using generative AI as '**cognitive co-pilots**' across the **direction, collection, processing, and dissemination** stages of the intelligence cycle could alleviate traditional challenges regarding the 'fact-poor and opinion-rich' environment analysts operate within. Nonetheless, **careful deployment involving frequent human validation** is crucial at this early stage of maturity and familiarity.

Autonomous agents – artificial entities that can sense their environment, make decisions, and take actions without human intervention – **could be an accelerating force within the**

intelligence and security context, due to their ability to draw on other data sources for additional validation. In theory, teams of agents could be used to rapidly process vast amounts of open-source data, provide preliminary risk assessments, and generate hypotheses for human analysts to explore further. However, until the underlying LLMs can provide reliable (consistent, correct, and safe) and accurate responses, agents will be at **risk of delivering unpredictable or misaligned outcomes**. The key mitigations in addressing these challenges are **accountability, transparency, and human oversight** of both the actions taken by the agent and the inference performed by the system.

To respond to the complex landscape outlined above, governments must devise policy interventions which have three main goals: **to create better visibility and understanding of generative AI systems; to promote best practices; and to establish incentives and enforcement of regulation**. Establishing signalling and reporting mechanisms into government and relevant third-party actors, and red lines in the highest-risk contexts (such as decision-making within critical national infrastructure) are important aspects of achieving these goals.

Signalling		Reporting		Red lines
<u>Watermarking</u> Automatically adding labels or invisible watermarks to AI-generated content is a possible technical solution to the challenges of generative AI-enabled disinformation . However, concerns persist over its vulnerability to deliberate tampering and the ability of bad-faith actors to bypass it entirely.	<u>Disclosure and explainability</u> The challenges associated with AI detection tools place additional emphasis on disclosing when generative AI is being used, and issuing clear guidance on appropriate use and warnings for misuse. Better outcomes will be co-dependent on the level of explainability provided by the system and individuals' ability to interpret AI outputs .	<u>Multi-layered and socio-technical evaluation</u> To understand the full spectrum of national security implications, AI system evaluation must go beyond the capabilities of any individual model. A multi-layered, socio-technical approach to system evaluation is needed to understand how human interactions and systemic factors interact with technical components of generative models to amplify different types of risks.	<u>Release strategies</u> Rapid increases in capability can mean policymakers are ill-prepared for the next game changing innovation. Leading AI developers recently committed to avoid releasing models without prior testing by government agencies, but this process must be open and transparent enough to ensure public trust in its conclusions.	Pre-empting the high-stakes contexts where generative AI should not be used will prevent situations where the technology can take irreversible actions without direct human oversight or authorisation .

Writing of this report coincided with the UK's AI Safety Summit in November 2023, and contemporaneous announcement of a new Government-sponsored AI Safety Institute for safety testing of the most advanced AI models. The coming months will be crucial in determining the role and scope of the new AI Safety Institute, and the UK's approach to managing emerging AI risks more broadly.

At the international level, there are two key actions the UK could take to narrow existing disparities between governance models: **promoting shared evaluation tools and clear targets; and contributing to international regulatory expertise and capacity**. The announcement of the new AI Safety Institute is a positive step in this regard, but the UK must make leaps in the three core areas of **compute, data, and staff** to meaningfully lead in this effort. Research into trustworthy LLMs inherently requires experts from different disciplines, including linguistics, computer science, cognitive psychology, cybersecurity, and policy.

Finally, achieving these global governance goals entails a minimum level of diplomatic engagement, which ensures that **rapid AI adoption does not supersede AI safety research**. Countries wishing to show leadership in AI safety must avoid undermining that positive work by allowing the fear of 'falling behind' adversaries to drive a race-to-the-bottom through high-risk applications.

Recommendations

AI system evaluation

- Building on the positive momentum from the AI Safety Summit, there are immediate steps the new AI Safety Institute should take to develop a world-leading AI evaluation ecosystem:
 - Prioritise a multi-layered, socio-technical approach to system evaluation so that novel system characteristics are scrutinised in addition to governance and application procedures.
 - Create a centralised register for generative AI model and system cards, which allows decision-makers across departments to review system details and make informed judgments about their risk appetite and applicability to envisaged use cases.

Intelligence analysis

- If generative AI is to be deployed operationally by the UK national security community, those organisations must ensure that user interfaces are designed to include explicit warnings about the accuracy and reliability of outputs, thus minimising the risks associated with over-trust or over-reliance.
- Additionally, detailed consideration should be given to how the use of LLMs in the national security context may affect warranty and legal compliance. The scale and opacity of LLMs means that purging information from them may be more challenging than for existing databases – targeting research resources at developing techniques such as ‘machine unlearning’ may help in addressing this challenge.

Autonomous agents

- LLM-augmented agent-based systems commissioned to perform autonomous actions should abide by certain requirements. The UK national security community must ensure that these requirements are met internally and should work through industry partners and trusted open-source community networks to encourage the same in those sectors:
 - Comply with frameworks such as the Open Worldwide Application Security Project (OWASP) design considerations to manage the risks of ‘Excessive Autonomy’. ‘Human-in-loop’ functionality must be included in these use cases.
 - Record actions taken and decisions made by the agents. The agent architecture must not obscure or undermine any potential aspects of explainability originating from the LLM.
 - Document what the agent-based system could do in a worst-case scenario.
 - Display warnings and caveats pertaining to the use of LLM generated output, at every stage of its commissioning, development, and deployment.

Cybersecurity and training

- The National Cyber Security Centre and Cabinet Office should develop guidance for safe generative AI use across government, which is aligned to users' proficiency and the cybersecurity risks within applications. For example, for experienced developers, AI code generation presents noteworthy efficiency gains and experimentation should be encouraged with appropriate validation techniques. For users less familiar with secure engineering practices, awareness training on the limitations and scrutiny of AI-generated code is essential.
- To encourage understanding of benefits and responsible use, departments should appoint liaisons to organise technical sessions where users can work with generative AI applications in sandboxed environments.

Disinformation and elections

- The Electoral Commission should partner with the Office of Communications (Ofcom) to develop new electoral rules for political parties' use of generative AI in the lead-up to the upcoming Parliamentary elections. These should demarcate generative uses which should be officially documented with the Electoral Commission.
- Ofcom's efforts should focus on public education campaigns to inform people of the ease with which generative AI can make convincing representations of high-profile political figures.

Radicalisation and terrorism

- The Home Office and Counter Terrorism Policing should commission research aimed at developing a more rigorous evidence base for terrorist uses of generative AI. A more detailed framework is needed to understand the stages of the radicalisation and recruitment lifecycle where generative AI may be leveraged.

Voice cloning

- The UK national security community should support a joint industry-academia initiative to address technical challenges in the voice cloning domain. This grouping should organise workshops and roundtables to gather leading audio specialists across academia, industry, and government to provide an assessment of the state-of-the-art in voice mimicry across accents and languages. This may lead to the establishment of a working group tasked with developing rigorous evaluation metrics for voice mimicry performance and detection.

Biochemical weapons

- The UK Biological Security Strategy proposed the development of a *National Biosurveillance Network* which would include a real-time *Bio Threats* radar to monitor threats and risks. Generative AI should be retrospectively included in this

monitoring framework.¹ Status reports and briefings should also be shared with the UK's Chemical Weapon Convention National Authority advisory committee in relation to chemical weapons applications and technologies.

CSAM

- The Home Office should issue clearer instruction on the legal status of *models* that have been trained on CSAM and of people who exchange model files without exchanging individual pieces of content. Guidance is also needed regarding what qualifies as illegal use of a generative AI system even if it has not been explicitly trained on CSAM.
- UK law enforcement agencies – led by the National Crime Agency – should coordinate with INTERPOL to create a new database of models used to generate CSAM. This would complement the existing Child Abuse Image Database (CAID). This could be a platform for exploring the creation of automated detection capabilities to detect when those models are used by criminals.

¹ Despite no direct references to generative AI, references are made to 'new, potentially extreme risks', new 'cyberbiosecurity risks' and broader misuse whereby 'more people now have the necessary skills to perform high risk research at low cost'. See <https://www.gov.uk/government/publications/uk-biological-security-strategy/uk-biological-security-strategy-html>.

1. An Introduction to Generative AI

Generative artificial intelligence (AI) is a form of AI that can generate content such as images, audio, and text based on user suggestions. These suggestions, or *prompts*, can take different forms: they might be a sketch image, a sample of audio such as a voice recording or a textual description of what to generate or summarise. Well-known examples of generative AI include *DALL-E* (OpenAI)², *Midjourney*³, and *Stable Diffusion* for generating images from text prompts; and *Bard* (Google)⁴, *ChatGPT* (OpenAI)⁵, and *LLaMA* (Meta AI)⁶ for generating text from text prompts.

1.1 A short history of AI

The sub-field of generative AI has emerged through decades of experimentation and iteration in the AI field, and this context needs to be understood to appreciate the origins of where we are today.

² OpenAI, Dall-E, <https://openai.com/research/dall-e>.

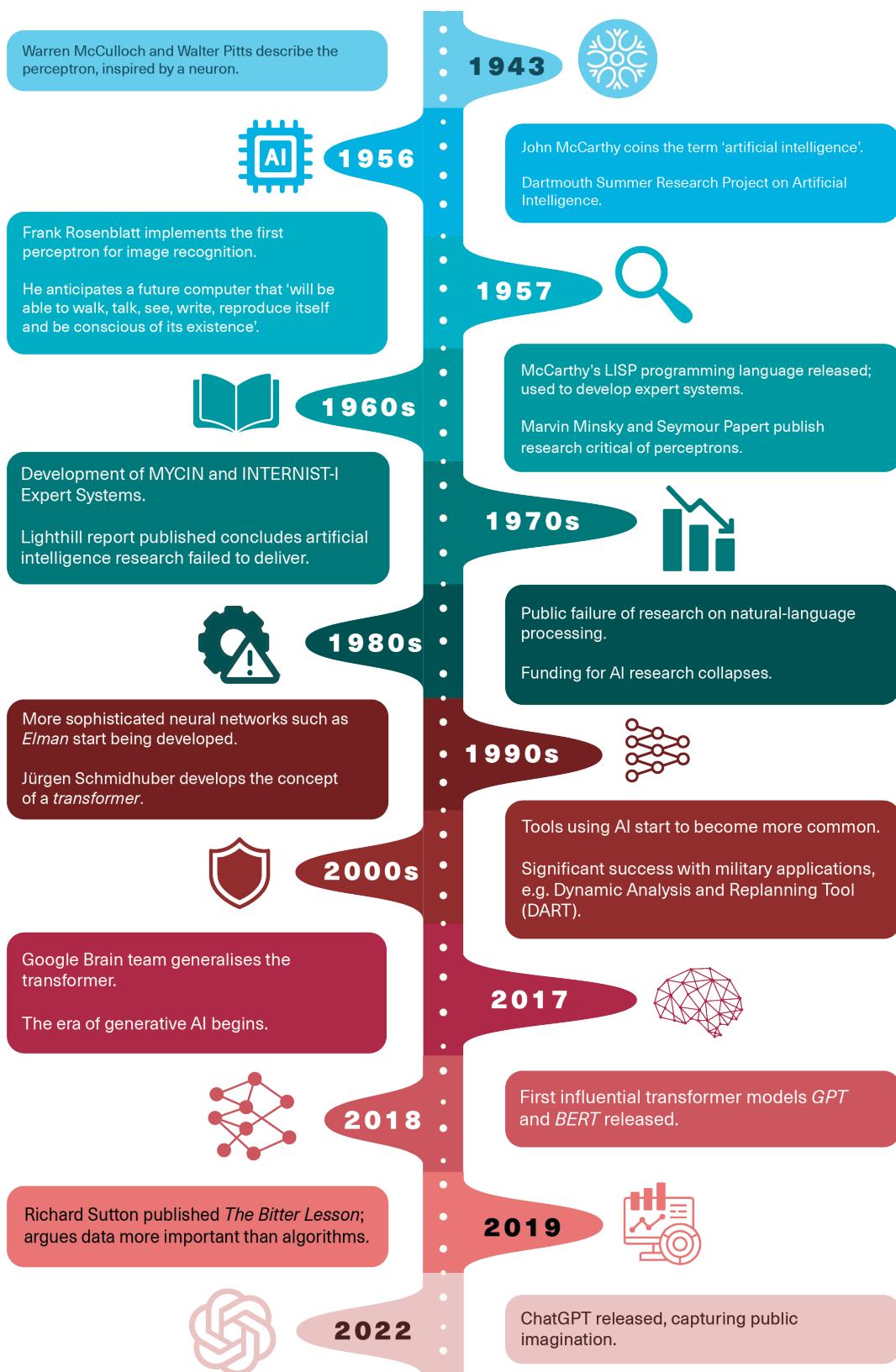
³ Midjourney, <https://www.midjourney.com>.

⁴ Google, Bard, <https://bard.google.com>.

⁵ OpenAI, ChatGPT, <https://openai.com/chatgpt>.

⁶ Meta, “Introducing LLaMA: A foundational 65-billion-parameter large language model,” published 24 February 2023, <https://ai.meta.com/blog/large-language-model-llama-meta-ai/>.

Figure 1. Timeline of generative of AI capturing the most significant moments in this story⁷



⁷ For a more detailed timeline, please see <https://cetas.turing.ac.uk/publications/rapid-rise-generative-ai>.

1.2 The pace of change

As demonstrated by the timeline, the history of AI is littered with hype and over-promises: from its first public failure at natural language processing in the 1960s, through to Microsoft's 'sexist' and 'racist' chatbot *Tay* in 2016⁸, AI has developed a reputation for under-delivering. The recent explosion of interest in generative AI is viewed cynically by some as a continuation of this pattern. But there are many ways in which generative AI represents a step-change in what is possible using AI. Previously, an expert team would create a specific task-based tool (such as route mapping or spell checking) for non-expert users, and they could effectively set boundaries for where users deployed that tool. With generative AI, users have far more latitude over how it is used, resulting in applications which tool creators would never have conceived of.⁹

As per Figure 1, the first influential transformer models emerged in 2017. These included the LLMs¹⁰ GPT ('Generative Pre-trained Transformer', OpenAI)¹¹ and BERT ('Bidirectional Encoder Representations from Transformers', Google).¹² Both GPT and BERT used a similar approach: a pretraining stage on a large corpus of data, which generates a general-purpose model, followed by task-specific fine tuning. This approach allows the model to be applied to a wide range of tasks without incurring considerable training costs. The general-purpose model is referred to as a *foundation model*.¹³

The following table illustrates the dramatic increase in the number of parameters and the number of tokens¹⁴ used to train LLMs.¹⁵

⁸ Elle Hunt, "Tay, Microsoft's AI chatbot, gets a crash course in racism from Twitter," *The Guardian*, 24 March 2016, <https://www.theguardian.com/technology/2016/mar/24/tay-microsofts-ai-chatbot-gets-a-crash-course-in-racism-from-twitter>.

⁹ Cem Dilmegani, "Top 100+ Generative AI Applications / Use Cases in 2023," *AIMultiple*, 26 October 2023, <https://research.aimultiple.com/generative-ai-applications/>.

¹⁰ Michael R Douglas, "Large Language Models," *arXiv* (October 2023), <https://arxiv.org/abs/2307.05782>.

¹¹ OpenAI, "Improving language understanding with supervised learning," 11 June 2018, <https://openai.com/research/language-unsupervised>.

¹² Jacob Devlin and Ming-Wei Chang, "Open sourcing BERT: State-of-the-Art Pre-training for Natural Language Processing," Google Research Blog, 2 November 2018, <https://blog.research.google/2018/11/open-sourcing-bert-state-of-art-pre.html>.

¹³ Rishi Bommasani et al., "On the opportunities and risks of foundation models," *arXiv* (August 2021), <https://arxiv.org/abs/2108.07258>.

¹⁴ Consider a *token* as an 'atom' of data, analogous to a syllable in language.

¹⁵ Daniel Gutierrez, "Introduction to GPT-3," Open Data Science, 25 August 2020, <https://opendatascience.com/introduction-to-gpt-3/>; Xavier Amatriain et al., "Transformer models: an introduction and catalog," *arXiv* (May 2023), <https://arxiv.org/abs/2302.07730>; Mohammed Lubbad, "GPT-4 Parameters: Unlimited guide NLP's Game-Changer," Mohammed Lubbad Medium, 19 March 2023, <https://medium.com/@mlubbad/the-ultimate-guide-to-gpt-4-parameters-everything-you-need-to-know-about-nlps-game-changer-109b8767855a>.

Year (Release)	Model	#Parameters	#Tokens
2018	GPT	110 million	1 billion
2018	BERT	340 million	3 billion
2019	GPT-2	1.5 billion	10 billion
2020	GPT-3	175 billion	500 billion
2022	PaLM	540 billion	780 billion
2023	GPT-4	1.8 trillion (estimated)	13 trillion

The sophistication of the models is non-linear. As the number of parameters grows and the size of the training dataset increases, LLMs frequently exhibit new properties (labelled as ‘emergent’).¹⁶ However, the way that models use parameters has evolved over time, meaning the number of parameters only provides a crude estimate of a model’s capabilities. For example, GPT-4 is a mixture of expert models, resembling several mid-sized models linked together rather than a single vast network.¹⁷

Developing an LLM from scratch – as opposed to fine-tuning a pre-trained model with all the security uncertainties¹⁸ and data poisoning risks¹⁹ this brings – is currently the preserve of the most well-funded organisations. As a result, the open-source LLM community, led by Hugging Face and Replicate, has expanded substantially since 2021, with more fine-tuned models released weekly.

This dramatic growth²⁰ is illustrated in the non-exhaustive table below. The resulting models are easier for a casual programmer to download and use – what was once the domain of a specialist is now accessible to anyone with basic knowledge of Python. In 2023, a leaked internal Google document claimed that open-source AI – driven by the February 2023 leak of LLaMA, a LLM developed by Meta – will outcompete Google and OpenAI, stating ‘we [Google] have no moat, and neither does OpenAI.’²¹ Competition from the open-source community is driving companies such as OpenAI to reverse their open policy, leading to

¹⁶ Jason Wei et al., “Emergent Abilities of Large Language Models,” *arXiv* (October 2022), <https://arxiv.org/abs/2206.07682>.

¹⁷ Maximilian Schreiner, “GPT-4 architecture, datasets, costs, and more leaked,” *The Decoder*, 11 July 2023, <https://the-decoder.com/gpt-4-architecture-datasets-costs-and-more-leaked>.

¹⁸ NCSC, “Thinking about the security of AI systems,” August 30, 2023, <https://www.ncsc.gov.uk/blog-post/thinking-about-security-ai-systems>.

¹⁹ NCSC, “Exercise caution when building LLMs,” NCSC Blog Post, 30 August 2023, <https://www.ncsc.gov.uk/blog-post/exercise-caution-building-off-langs>.

²⁰ Suleiman Kazi and Adel Elmahdy, “Top Large Language Models (LLMs): GPT-4, LLaMA 2, Mistral 7B, ChatGPT, and More,” Vectara Blog Post, 17 October 2023, <https://vectara.com/top-large-language-models-llms-gpt-4-llama-gato-bloom-and-when-to-choose-one-over-the-other>.

²¹ Dylan Patel and Afzhal Ahmad, “Google ‘We have no Moat, and Neither Does OpenAI’: Leaked Internal Google Document Claims Open Source AI Will Outcompete Google and OpenAI”, Semianalysis, 4 May 2023, <https://www.semianalysis.com/p/google-we-have-no-moat-and-neither>.

fears that the next leaps forward will happen behind closed doors.²² Longer-term, LLMs may undergo more incremental efficiency gains: models will be smaller, with less data needed for fine-tuning, while being cheaper to run and more environmentally friendly.²³

Year	Model	Creator	Notes
2022	BLOOM ²⁴	BigScience	Collaboration of over 1000 researchers from over 250 institutions
2022	FLAN UL2	Google	Apache-2.0 license allowing commercial use
2023	LLaMA ²⁵	Meta AI	Available for academic use on application to Meta
2023	Alpaca ²⁶	Stanford	Fine-tuned from LLaMA; not available for commercial use
2023	ChatGLM ²⁷	Tsinghua University	Chinese/English LLM; Apache license allowing commercial use
2023	LLaMA 2 ²⁸	Meta AI	Free for research and commercial use
2023	Claude 2 ²⁹	Anthropic	Currently only available in the US and UK
2023	MPT-7B ³⁰	MosaicML	Open-source; licenced for commercial use
2023	Falcon LLM ³¹	TII	Free for research and commercial use
2023	Persimmon-8B ³²	Adept	Open-source; Apache license allowing commercial use
2023	Vicuna-13B ³³	LMSYS Org	Open-source; licenced for non-commercial use only
2023	Mistral 7B ³⁴	Mistral AI	Apache-2.0 license allowing commercial use
2023	Dolly 2.0 ³⁵	Databricks	Open-source; licenced for commercial use

²² Will Douglas Heaven, “The open-source AI boom is built on Big Tech’s handouts. How long will it last?,” *MIT Technology Review*, 12 May 2023, <https://www.technologyreview.com/2023/05/12/1072950/open-source-ai-google-openai-eleuther-meta/>.

²³ Tianyi Chen et al., “LoRAShear: Efficient large language model structure pruning and knowledge recovery,” *arXiv* (October 2023), <https://arxiv.org/abs/2310.18356>.

²⁴ Teven Le Scao et al., “BLOOM: a 176B-parameter open-access multilingual language model,” *arXiv* (June 2023), <https://arxiv.org/abs/2211.05100>.

²⁵ Hugo Touvron et al., “LLaMA: Open and efficient foundation language models,” *arXiv* (February 2023), <https://arxiv.org/abs/2302.13971>.

²⁶ Rohan Taori et al., “Alpaca: A Strong, Replicable Instruction-Following Model,” Stanford University, 13 March 2023, <https://crfm.stanford.edu/2023/03/13/alpaca.html>.

²⁷ ChatGLM, <https://chatglm.cn>.

²⁸ Meta, LLaMA, <https://ai.meta.com/llama/>.

²⁹ Anthropic, Claude 2, <https://www.anthropic.com/index/claude-2>.

³⁰ MosaicML NLP team, “Introducing MPT-7B: A New Standard for Open-Source, Commercially Useable LLMs,” MosaicML Research Blog, 5 May 2023, <https://www.mosaicml.com/blog/mpt-7b>.

³¹ Technology Innovation Institute (TII), Falcon 180B, <https://falconlm.tii.ae>.

³² Erich Elsen et al., “Releasing Persimmon-8B,” Adept Blog, 7 September 2023, <https://www.addept.ai/blog/persimmon-8b>.

³³ Vicuna, “Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality,” LMSYS Org Blog, 30 March 2023, <https://lmsys.org/blog/2023-03-30-vicuna/>.

³⁴ Mistral AI team, “Mistral 7B,” Mistral AI News, 27 September 2023, <https://mistral.ai/news/announcing-mistral-7b/>.

³⁵ Mike Conover et al., “Free Dolly: Introducing the World’s First Truly Open Instruction-Tuned LLM,” Data Bricks Blog, 12 April 2023, <https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-lm>.

Despite rapid improvements in performance, for many observers LLMs have become infamous for their 'hallucinations'.³⁶ These 'hallucinations' can lead to a general lack of trust in the technology³⁷ and even lawsuits.³⁸ When Google's Bard 'hallucinated' during its first public demonstration, Alphabet briefly lost \$100 billion in market value.³⁹

Such hallucinations exemplify how LLMs can blur the boundary between real and fake, reliable and unreliable. The last year has seen the use of AI in day-to-day life transition from predominantly spell-checking to producing sonnets through ChatGPT or art through DALL-E. As AI systems become more refined, it may become impossible to detect whether the text in those sonnets was generated by humans. The blending or layering of real and fake content only makes this more challenging. This ambiguity has the potential to further degrade institutional trust around the world.⁴⁰

The increasingly diverse range of AI applications has been made possible by the growth in computational power and access to ever larger datasets via the internet. According to OpenAI, the amount of computational power used to train the largest AI models has doubled every 3.4 months since 2012.⁴¹ Richard Sutton argues in his influential 2019 essay, *The Bitter Lesson*,⁴² that the availability of more data has played a far greater role than improvements to the underlying neural network architectures and algorithms that train them. Data quality can also significantly influence the success of a model;⁴³ if large datasets come at the cost of introducing low-quality data, the rate of progress might slow. Access to

³⁶ Robin Emsley, "ChatGPT: these are not hallucinations – they're fabrications and falsifications," *Schizophrenia* 9, no. 52 (August 2023), <https://www.nature.com/articles/s41537-023-00379-4>; Nick McKenna, "Sources of hallucinations by large language models on inference tasks," *arXiv* (October 2023), <https://arxiv.org/abs/2305.14552>.

³⁷ Avishek Choudhury and Hamid Shamszare, "Investigating the Impact of User Trust on the Adoption and Use of ChatGPT: Survey Analysis," *Journal of Medical Internet Research* 25, no. 1 (2023).

³⁸ Sabrina Ortiz, "ChatGPT's hallucination just got OpenAI sued. Here's what happened," *ZDNET*, 9 June 2023, <https://www.zdnet.com/article/chatgpts-hallucination-just-got-openai-sued-heres-what-happened>.

³⁹ Emily Olsen, "Google shares drop \$100 billion after its new chatbot makes a mistake," *NPR*, 9 February 2023, <https://www.npr.org/2023/02/09/1155650909/google-chatbot--error-bard-shares>.

⁴⁰ Blayne Haggart, "Here's why ChatGPT raises issues of trust," *World Economic Forum*, 6 February 2023, <https://www.weforum.org/agenda/2023/02/why-chatgpt-raises-issues-of-trust-ai-science>.

⁴¹ Karen Hao, "The computing power needed to train AI is now seven times faster than ever before," *MIT Technology Review*, 11 November 2019, <https://www.technologyreview.com/2019/11/11/132004/the-computing-power-needed-to-train-ai-is-now-rising-seven-times-faster-than-ever-before>.

⁴² Rich Sutton, "The Bitter Lesson," Incomplete Ideas (personal website), 13 March 2019, <http://www.incompleteideas.net/IncompleteIdeas/BitterLesson.html>.

⁴³ Tom Brown et al., "Language models are few-shot learners," in *Advances in Neural Information Processing Systems* 33 (Neurips 2020), 1877-901, https://proceedings.neurips.cc/paper_files/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

high-quality data might also be reduced by new technologies and legislation, as organisations and individuals seek to protect copyrighted material.⁴⁴

The potential economic impact of generative AI is difficult to quantify. OpenAI has estimated that ‘around 80% of the US workforce could have at least 10% of their work tasks affected by the introduction of LLMs’.⁴⁵ Similar forecasts have been made before with new technological breakthroughs, but the outcome tends to be more nuanced. For example, countries with the highest rates of automation and robotics – such as Japan (264 robots per 10,000 employees) – tend to have the lowest unemployment.⁴⁶ Recent estimates claim that generative AI could add \$2.6 trillion – \$4.4 trillion annually to the global economy across 63 use cases (the UK’s entire GDP in 2021 was \$3.1 trillion),⁴⁷ but questions remain as to how such additional GDP would be distributed across populations.

1.3 Methodology

This study sought to address the following four research questions:

- **RQ1:** What social, political, and security risks are presented by the widespread use of generative AI models, with particular focus on generative language models?
- **RQ2:** What is needed in terms of technical and policy requirements to be able to identify and analyse synthetically generated media and reliably distinguish it from human-generated media?
- **RQ3:** Which stages of the AI supply chain should be prioritised to create safeguards which adequately prevent the misuse of generative AI tools, and what additional policy, guidance, or training is required to this effect?
- **RQ4:** What domestic and international policy and regulatory options are available to respond to the potential risks posed by the proliferation of generative AI tools (identified in RQ1)?

⁴⁴ Nicola Lucchi, “ChatGPT: A Case Study on Copyright Challenges for Generative Artificial Intelligence Systems,” *European Journal of Risk Regulation* (August 2023), 1-23, <https://doi.org/10.1017/err.2023.59>; Karen Hao, “The computing power needed to train AI is now seven times faster than ever before,” *MIT Technology Review*, 11 November 2019, <https://www.technologyreview.com/2019/11/11/132004/the-computing-power-needed-to-train-ai-is-now-rising-seven-times-faster-than-ever-before/>; Sneha Singh, “BBC takes measures to restrict ChatGPT AI from crawling its content,” *Techstory*, 9 October 2023, <https://techstory.in/bbc-takes-measures-to-restrict-chatgpt-ai-from-crawling-its-content>.

⁴⁵ Tyna Eloundou et al., “GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models,” *arXiv* (March 2023), <https://arxiv.org/abs/2303.10130>.

⁴⁶ International Federation of Robotics, “Robot Race: The World’s Top 10 automated countries,” IFR Press Release, 27 January 2021, <https://ifr.org/ifr-press-releases/news/robot-race-the-worlds-top-10-automated-countries>.

⁴⁷ Michael Chui et al., “The economic potential of generative AI: the next productivity frontier,” McKinsey Report, 14 June 2023, <https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier#business-value>.

To this end, the project team conducted semi-structured interviews and focus groups between June and September 2023 with 50 participants across academia, civil society organisations, government, and industry. These participants were identified through a purposive sampling strategy to ensure informed responses to the research questions. A snowball sampling method enabled the identification of further suitable participants for interview. A semi-structured interview approach meant the line of questioning across interviews was consistent while allowing for elaboration in response to a participant's specific area of expertise.

Following the conclusion of interviews, notes were analysed through a general inductive approach whereby meaning is extracted from data and categorised into relevant themes and sub-themes. Interviews were conducted on an anonymised basis.

The findings are also informed by a closed, invitation-only workshop held by CETaS in October 2023 titled, "Large Language Models and Terrorism: Legal and Policy Considerations". This half-day session gathered experts from academia, industry, civil society, law enforcement and government, and is referenced primarily in the sub-section on "Radicalisation and Terrorism" in Chapter 2.

The research team conducted a targeted literature review to map key developments in the field of generative AI over time; social and political risks posed by the technology; opportunities and limitations in the intelligence context; unanswered technical questions or challenges in the field; and the range of governance and policy responses available to policymakers.

The technical component of this research project involved the incorporation of information from two distinct projects (see Case Studies 1 and 2 at the end of the report), each demonstrating different aspects of the role of language agents. The first project – commissioned directly for this report – explored the application of language agents in open-source intelligence. This entailed a review of projects on GitHub, with selection made based on a predetermined set of criteria. The chosen project, LLM_OSINT, underwent a thorough evaluation with detailed notes on system use and test runs being comprehensively documented. Conversely, the second case study leveraged an existing research project, named Gen-MAS-Sim, which aimed at employing language agents to simulate human behaviours. Despite not being originally devised to support this report, it was included due to its relevance and overlap with the research theme. Supplemental analysis was performed which included an evaluation of Gen-MAS-Sim's performance and limitations.

One limitation of this project is the lack of dedicated legal expertise within the project team, which resulted in a more restricted analysis of the legal status of generative AI and LLMs. This is an important avenue for future research.

Research participants took part in this study in a personal capacity. The views and responses expressed here reflect participant opinions and should not be interpreted to represent the official position of any government department, agency, or other organisation.

2. Evaluating Political, Digital and Physical Security Risks

The pace of change described in Chapter 1 has caused concerns about the nature of political, digital, and physical risks posed by generative AI. It has also led some to ask whether developments in generative AI are better understood as **augmenting pre-existing societal risks** or as posing **completely novel risks**. In one sense, this formulation is useful in developing a clearer timeline of generative AI risks. However, in the context of political disinformation and influence operations, the increase in speed and scale offered by generative AI to malicious actors raises the exposure of a larger proportion of the population than before.⁴⁸

In Chapter 4 of this report, we will describe a multi-layered, socio-technical framework to evaluate national security risks from generative AI. But before evaluating those risks, a rigorous breakdown is required of where they sit within the broader security landscape and how *malicious* generative AI use cases differ from *incidental* sources of risk.

As outlined in Brundage et al. (2018), malicious AI uses consist of threats to the following domains:⁴⁹

- *Political* security: the use of AI to automate tasks pertaining to surveillance, persuasion, and deception as well as novel attacks that take advantage of an improved capacity to analyse human behaviours, moods, and beliefs based on available data.
- *Digital* security: the use of AI to automate tasks pertaining to cyberattacks as well as novel attacks that exploit human vulnerabilities, existing software vulnerabilities or the vulnerabilities of AI systems themselves.
- *Physical* security: the use of AI to automate tasks pertaining to attacks on physical systems as well as novel attacks that subvert cyber-physical systems or involve physical systems that would be infeasible to direct remotely.

⁴⁸ Interview with industry representative, 25 July 2023.

⁴⁹ Miles Brundage et al., “The Malicious Use of Artificial Intelligence: Forecasting, Prevention and Mitigation,” Future of Humanity Institute, University of Oxford & Centre for the Study of Existential Risk, University of Cambridge, Center for a New American Security, Electronic Frontier Foundation, and OpenAI Report, 20 February 2018, <https://img1.wsimg.com/blobby/go/3d82daa4-97fe-4096-9c6b-376b92c619de/downloads/MaliciousUseofAI.pdf?ver=1553030594217>.

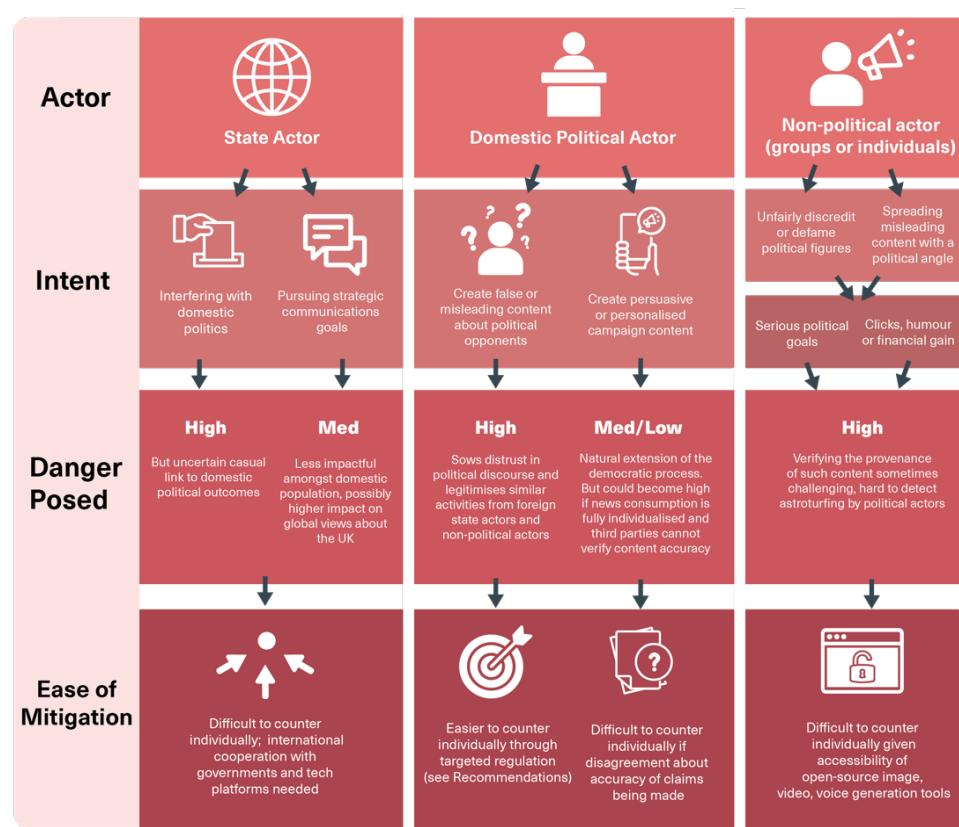
2.1 Political security

2.1.1 Political disinformation and electoral interference

Across research interviews for this project, disinformation was the most referenced generative AI risk.⁵⁰ This analysis focuses predominantly on political disinformation; other types of disinformation may carry different considerations, but the way that generative AI could act as a **force multiplier** in upcoming democratic elections merits additional academic scrutiny.

The diagram below contextualises the subsequent discussion of the role of generative AI in the political information ecosystem: outlining the different actors, their intent, the level of danger posed and the ease of mitigation against undesirable uses of generative AI in this context.

Figure 2. The role of generative AI in the political information ecosystem



⁵⁰ Findings from a CETaS workshop in October 2022 highlighted that AI technology (such as Generative Pre-Trained Transformer 3, GPT-3) is starting to generate more convincing and realistic narratives at speed. Combined with enhanced machine translation capabilities, this could enable automated generation of false news and information at a large scale.

Generative AI can leverage various modalities to make the task of distinguishing between ‘real’ and ‘fake’ extremely challenging. The **cumulative effect of generative text, image, video and audio** combined as part of a larger influence operation will exceed the impact that any one of those modalities can have individually.⁵¹ For example, an AI-generated video of a prominent politician delivering a speech at a venue they never attended may be seen as more plausible if presented with an accompanying stack of audio and imagery, such as the politician taking questions from reporters paired with text-based journalistic articles covering the content of the supposed speech. One interviewee distinguished between interactional and compositional deepfakes: ‘interactional deepfakes refer to multimodal content that engage with users, which could represent a huge leap in immersion. Compositional deepfakes are where users may create false histories of targets to discredit them with synthetic videos or images. When this content is spread, it becomes so difficult to discern what is slander or not and creates huge risks in breaking down trust.’⁵²

An alternative approach may be to blend genuine images with disingenuous video or audio. The undermining of existing communication and evidence-based mechanisms could be as significant as the ability to persuade people of falsities.⁵³ The invention of sources could cast doubt on whether citations can be trusted as a meaningful signal of authority and potentially fuel conspiracy theories.⁵⁴ One interviewee posited that if the nature of the Internet is such that ‘you can have one truth but an infinite number of lies, what are the chances of a chatbot spreading misinformation when it does not know what “truth” is?’⁵⁵

At a quantitative level, the **scale** of information operations could be significantly enhanced by improvements in usability, reliability, and cost-effectiveness of LLMs. There tends to be a limited range of fixed narratives that disinformation actors seek to perpetuate, so having LLMs try to produce hundreds of *new* narratives will be of limited utility. However, for those pre-defined narratives, they will be crucial in generating masses of content which supports their dissemination. At a qualitative level, the **personalisation** of these operations could reach new levels of convincingness with more impressive storytelling capabilities and individually tailored disinformation campaigns no longer facing the same resource

⁵¹ Interview with academic, 20 June 2023; Mustafa Suleyman, “Inflection AI co-founder Mustafa Suleyman: ‘Ban the use of AI in elections – right now’,” *Fortune*, 5 September 2023, <https://fortune.com/2023/09/05/inflection-ai-co-founder-mustafa-suleyman-ban-use-of-ai-elections-right-now-artificial-intelligence-2024-politics>.

⁵² Interview with academic, 11 July 2023; Eric Horvitz, “On the Horizon: Interactive and Compositional Deepfakes,” in *Proceedings of the 2022 International Conference on Multimodal Interaction* (ICMI ‘22), 653–661, <https://dl.acm.org/doi/abs/10.1145/3536221.3558175>.

⁵³ Interview with academic, 11 July 2023.

⁵⁴ Chris Moran, “ChatGPT is making up fake Guardian articles. Here’s how we are responding,” *The Guardian*, 6 April 2023, <https://www.theguardian.com/commentisfree/2023/apr/06/ai-chatgpt-guardian-technology-risks-fake-article>.

⁵⁵ Interview with civil society representative, 7 July 2023; Di Cooke, “Synthetic Media and Election Integrity: Defending our Democracies,” *CETaS Expert Analysis* (August 2023), <https://cetas.turing.ac.uk/publications/synthetic-media-and-election-integrity-defending-our-democracies>.

constraints.⁵⁶ Audiences may be targeted through auto-generated persuasive messages at scale, while also being targeted via one-to-one messaging-based campaigns.⁵⁷ This could involve propagandists investing in fine-tuning LLMs by incorporating bespoke data (such as user engagement data) that increases resonance with intended targets.⁵⁸

While some interviewees cautioned about a lack of robust empirical data on malicious actors using generative AI to sway individuals or communities, there are signs that it may represent a natural methodological progression in the **electoral landscape**.⁵⁹ For example, in the run up to the September 2023 Slovakian parliamentary elections, videos featuring AI-generated voices of politicians spread across social media and messaging platforms.⁶⁰ Coordinated campaigns going live in the hours or days preceding voting (as in the Slovakian case) are particularly concerning because of the length of time it can take factcheckers to identify an issue and provide a rebuttal.

Looking ahead to 2024, there is nervousness regarding upcoming elections in the UK, US, India, and European Parliament.⁶¹ In the UK and US, patterns are already emerging which are cause for concern. In October 2023, a fake audio recording of Keir Starmer MP, the leader of the Labour Party, was widely circulated on the first day of the Labour Party conference,⁶² while one interviewee described how members of US Congress were already being approached with ‘hyper-customised LLM campaign strategies’.⁶³ In the NCSC’s 2023 Annual Report, it assessed that ‘LLMs will almost certainly be used to generate fabricated content (...) and that deepfake campaigns are likely to become more advanced in the run up to the next nationwide vote’, also concluding that elections ‘almost certainly represent

⁵⁶ Interview with industry representative, 26 June 2023; Ben Buchanan, Andrew Lohn and Micah Musser, *Truth, lies, and automation: How language models could change disinformation* (CSET Georgetown: May 2021), <https://cset.georgetown.edu/publication/truth-lies-and-automation/>.

⁵⁷ Thor Benson, “This Disinformation Is Just for You,” *Wired*, 1 August 2023, <https://www.wired.com/story/generative-ai-custom-disinformation>; Percy Liang et al., “Holistic Evaluation of Language Models,” *arXiv* (October 2023), <https://arxiv.org/abs/2211.09110>.

⁵⁸ Interview with academic, 1 August 2023; Josh A. Goldstein et al., “Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations,” Georgetown University’s Center for Security and Emerging Technology, OpenAI and Stanford Internet Observatory Joint Report, <https://cyber.fsi.stanford.edu/io/publication/generative-language-models-and-automated-influence-operations-emerging-threats-and>.

⁵⁹ Interview with academic, 11 July 2023; Interview with government representative, 19 July 2023.

⁶⁰ Olivia Solon, “Trolls in Slovakian Election Tap AI Deepfakes to Spread Disinfo,” *Bloomberg*, 29 September 2023, <https://www.bloomberg.com/news/articles/2023-09-29/trolls-in-slovakian-election-tap-ai-deepfakes-to-spread-disinfo>.

⁶¹ 2024 will see 65 elections across 54 countries – there will not be this many again until 2048. See Katie Harbath and Ana Khizanishvili, “Insights from data: what the numbers tell us about elections and future of democracy,” Integrity Institute, 10 March 2023, <https://integrityinstitute.org/blog/insights-from-data>.

⁶² Morgan Meaker, “Deepfake Audio is a Political Nightmare,” *Wired*, 9 October 2023, <https://www.wired.co.uk/article/keir-starmer-deepfake-audio>.

⁶³ Interview with industry representative, 19 July 2023.

attractive targets for malicious actors and so organisations and individuals need to be prepared for threats, old and new'.⁶⁴

Given access to fine-grained data on minority communities from polls, data brokers or social media platforms, it will become possible to ‘develop content for a coherent persona, allowing propagandists to build credibility with a target audience without actually knowing that audience’.⁶⁵ Chatbots that use personal pronouns and emojis were highlighted as particularly interesting in this regard, feeding into the anthropomorphism already prevalent with these tools and leading people to believe they are conversing with something that ‘is on their side’.⁶⁶

Several papers have carried out studies to determine whether people are more easily deceived by AI or human-generated misinformation.⁶⁷ In experiments of GPT-3 capabilities,⁶⁸ human participants were able to distinguish multi-paragraph GPT-3 news articles from authentic news articles at a rate only slightly better than random chance⁶⁹ while a Stanford University study found that research participants became “significantly more supportive” of policies on smoking bans, gun control and carbon taxes when reading AI-produced texts.⁷⁰

⁶⁴ NCSC, “NCSC warns of enduring and significant threat to UK’s critical infrastructure,” NCSC News, 14 November 2023, <https://www.ncsc.gov.uk/news/ncsc-warns-enduring-significant-threat-to-uks-critical-infrastructure>.

⁶⁵ Josh A. Goldstein and Girish Sastry, “The Coming Age of AI-Powered Propaganda,” *Foreign Affairs*, 7 April 2023, <https://www.foreignaffairs.com/united-states/coming-age-ai-powered-propaganda>.

⁶⁶ Michael Atleson, “The Luring Test: AI and the engineering of consumer trust,” US Federal Trade Commission Business Blog, 1 May 2023, <https://www.ftc.gov/business-guidance/blog/2023/05/luring-test-ai-engineering-consumer-trust>. For example, Inflection’s chatbot – Pi – frequently uses emojis during its conversations, see <https://inflection.ai>.

⁶⁷ Giovanni Spitale, Nikola Biller-Andorno and Federico Germani, “AI model GPT-3 (dis) informs us better than humans,” *arXiv* (January 2023), <https://arxiv.org/abs/2301.11924>; Matthew Groh et al., “Deepfake detection by human crowds, machines, and machine-informed crowds,” *arXiv* (October 2021), <https://arxiv.org/abs/2105.06496>.

⁶⁸ According to OpenAI, GPT-4 is 40% more likely to produce factual content than GPT-3.5. However, an analysis by NewsGuard found that GPT-4 is more susceptible at generating misinformation (and in a more convincing manner) than GPT-3.5. Whereas GPT-3.5 refused to advance 20 of the 100 false narratives posed to it, GPT-4 generated all 100 falsehoods. See Lorenzo Arvanitis et al., “Despite OpenAI’s promises, the company’s new AI tool produces misinformation more frequently, and more persuasively, than its predecessor,” Misinformation Monitor: March 2023, NewsGuard, March 2023, <https://www.newsguardtech.com/misinformation-monitor/march-2023/>.

⁶⁹ Tom Brown et al., “Language models are few-shot learners,” in *Advances in Neural Information Processing Systems 33* (Neurips 2020), 1877-901, https://proceedings.neurips.cc/paper_files/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

⁷⁰ Hui Bai et al., “Artificial intelligence can persuade humans on political issues,” OSF PrePrints, 17 October 2023, <https://osf.io/stakv>.

It is important to stress that although disinformation was seen as an obvious opportunity for malicious actors, there was **scepticism** about whether generative AI would upend existing ways of operating. Hostile actors perpetuating disinformation must also be able to exploit the latest technology. For example, they may face constraints in training users to deploy models in ways that inflict the most damage, or they may lack the compute needed to achieve the scale required to influence an electoral process. Moreover, it is not a given that generative AI will be useful for the type of disinformation that they specialise in.⁷¹

One interviewee concurred by saying, ‘if a nation state wants to do a disinformation campaign, they do not need generative AI (...) it does not actually help you with the hard parts of a scalable disinformation campaign. You still need the infrastructure to get stuff out there and the means of getting it in the right spaces.’⁷² In this vein, it is important to recognise that generative AI may help in the production of false, misleading, and inauthentic content, but not necessarily its distribution.

Relatedly, it is not immediately clear that generative AI will make it inherently harder for governments to detect and mitigate disinformation. The work that is done to shut down information threats is not usually content specific but behavioural – for example, looking at associations between different accounts to give clues that something is amiss.^{73 74} Moreover, AI could play a role in detecting fake stories online by using natural language processing to help detect semantic features characteristic of fake news or analysing the patterns of news spread on social networks, which is typically shared differently to real news stories.⁷⁵ This suggests that the frameworks to understand approaches to disinformation campaigns may require tweaking rather than radical transformation.

2.1.2 Surveillance, monitoring and geopolitical fragmentation

It is important to move beyond a domestic focus to sufficiently understand the nature of the political security threat posed by generative AI. Globally, the number of democracies has

⁷¹ Interview with industry representative, 2 August 2023.

⁷² Interview with government representative, 18 August 2023.

⁷³ As per the often-cited ABC framework (actors, behaviour, content) for mis/disinformation campaigns, generative AI may make it cheaper and easier to produce more persuasive content, but it may not necessarily impact the actors or their behaviours to the same degree.

⁷⁴ Interview with academic, 1 August 2023; Josh A. Goldstein et al., “Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations,” Georgetown University’s Center for Security and Emerging Technology, OpenAI and Stanford Internet Observatory Joint Report, <https://cyber.fsi.stanford.edu/io/publication/generative-language-models-and-automated-influence-operations-emerging-threats-and>.

⁷⁵ OECD, “AI language models: Technological, socio-economic and policy considerations,” OECD Digital Economy Papers, No. 352, 13 April 2023, https://www.oecd-ilibrary.org/science-and-technology/ai-language-models_13d38f92-en.

started to decrease year-on-year⁷⁶ and authoritarian states' use of emerging technology can play a decisive role in perpetuating that trend. While concerns have mostly focused on the potential to boost propaganda efforts against the West, it should not be overlooked that democratic societies receive only a small fraction of the propaganda authoritarian countries distribute to their own populations. For example, in 2019, Xi Jinping ordered the leveraging of AI to 'comprehensively increase' the CCP's ability to mould public opinion.⁷⁷ Sir Richard Moore, the head of SIS, alluded to this theme in a recent speech⁷⁸:

'China benefits from sheer scale: AI, in its current form, requires colossal volumes of data; the more data you have, the more rapidly you can teach machine-learning tools. China has added to its immense datasets at home by hoovering up others abroad. And the Chinese authorities are not hugely troubled by questions of personal privacy or individual data security. They are focused on controlling information and preventing inconvenient truths from being revealed.'

Three core pieces of analysis emerge from Sir Richard Moore's diagnosis:

First, the context of the 'Digital Silk Road'. While China has watered down some of its infrastructure investments in the Belt and Road Initiative, the global proliferation of technology which adheres to Chinese **standards and values** continues apace. One interviewee described a Chinese version of an anime cartoon generator which grew in popularity across Latin America – yet its training data was extremely biased and therefore produced severe errors when generating faces of people of colour.⁷⁹ Where discriminatory technology proliferates in countries which may already face challenges to political stability, there are potentially significant repercussions for global security.

Second, the role of generative AI in shaping the collective memory of the Chinese internet and society. Compared to mediums like folk music and stories, generative AI models are easier to penetrate for the CCP given their significant influence within Chinese technology and innovation sectors. Consequently, such technologies could be extremely useful for their attempts to control depictions of history and enforce a **single version of truth** for future generations.⁸⁰ This effect is amplified as these models are increasingly used across

⁷⁶ Bastian Herre, "The world has recently become less democratic," Our World in Data article, 6 September 2022, <https://ourworldindata.org/less-democratic>.

⁷⁷ Bill Drexel and Caleb Withers, "Generative AI could be an authoritarian breakthrough in brainwashing," *The Hill*, 26 February 2023, <https://thehill.com/opinion/technology/3871841-generative-ai-could-be-an-authoritarian-breakthrough-in-brainwashing>.

⁷⁸ Sir Richard Moore, "Speech by Sir Richard Moore, Head of SIS, 19 July 2023," HMG, 19 July 2023, <https://www.gov.uk/government/speeches/speech-by-sir-richard-moore-head-of-sis-19-july-2023>.

⁷⁹ Interview with government representative, 18 August 2023.

⁸⁰ Ibid.

occupational or recreational domains and reinforce disproportionate surveillance infrastructures.

Third, there is a need to consider how improvements in the generative AI landscape relate to the **theoretical defence of democracy**. One experiment in technical innovation demonstrated that GPT-4 can come up with more creative ideas than rival human creators – if taken to its theoretical conclusion, this could eventually undermine the notion that freedom of expression makes democracies more economically and politically viable than autocracies.⁸¹ In more practical terms, the free and open nature of democracies means that those who wish to use the creative traits of generative AI for malign purposes have an easier time doing so against democracies than autocracies, where those traits are more easily stifled by the political system.

Questions of political theory are increasingly interlinked with the direction of emerging technologies. These questions are partly borne out of an understanding that there is a clearer sense of what needs protecting in the information space within authoritarian states than there is in democracies – this is a perverse situation when considering that a healthy and secure information ecosystem is essential to the proper functioning of democracies.⁸² In China, the techno-nationalist discourse around LLMs across traditional and social media reflects the more prominent role that the government plays in shaping the AI ecosystem (for example via significant compute funding) and ensuring that the interests of academia and industry align with those of the state.⁸³ In the words of one interviewee, this type of ‘lever pulling’ has a clear purpose:

‘What matters to the CCP is regime survival, regime survival and regime survival. So long as they have continued economic development which keeps the middle classes happy, that tacit agreement regarding privacy and human rights infringements continues. And a big part of that economic development is how well the technology sectors are doing.’⁸⁴

This emphasises the need to understand the **cultural and behavioural aspects to technology use** around the world.⁸⁵ Playing a leading role in the generative AI landscape will mean different things to different countries and applications will vary widely. Nonetheless, at least at the level of principles, there are some positive signs of global alignment. China’s attendance at the UK’s AI Safety Summit in November 2023 was a

⁸¹ Interview with industry representative, 31 July 2023.

⁸² Interview with industry representative, 31 July 2023.

⁸³ Interview with industry representative, 31 July 2023; Jeffrey Ding and Jenny Xiao, “Recent trends in China’s large language model landscape,” Centre for the Governance of AI Paper, 28 April 2023, <https://www.governance.ai/research-paper/recent-trends-chinas-ilm-landscape>.

⁸⁴ Interview with government representative, 18 August 2023.

⁸⁵ Ibid.

diplomatic coup, while their Interim Measures set out obligations regarding content management, protection and security of personal data, and transparency of generative AI in China.⁸⁶

More attention is devoted to global governance matters in Chapter 4, but it is important to stress how global approaches to generative AI development and implementation are directly linked to political security at home and abroad.

2.2 Digital security

2.2.1 Cybersecurity

In many cases, generative AI is an amplifier of pre-existing cybersecurity risks. By reducing the degree of specialist knowledge required, generative AI can **assist the less technically able user** in experimenting with novel cyberattack techniques and increase their sophistication iteratively to result in capable attacks.^{87 88} Less certain is whether generative AI will enable wholly new types of cyberattack that even the best cyberhackers would not have been aware of before, making them extremely difficult to combat. In the longer run, this will be the most significant concern from a national security perspective.

When considering model security, two growing areas of concern are the ability to **poison models** and the data they are trained on⁸⁹; and the ability to **manipulate, subvert or otherwise inject prompts** with malicious instructions.⁹⁰ Regarding the former, the size of today's LLMs make it impossible to know the totality of the data they contain. This helps potential attackers disguise the manipulation of very small quantities of data which nonetheless create insecurities.⁹¹ One May 2023 research paper showed that by using as

⁸⁶ Joshua Cole et al., "New generative AI measures in China," Ashurst Insights, 26 September 2023, <https://www.ashurst.com/en/insights/new-generative-ai-measures-in-china/>.

⁸⁷ Although there are reservations regarding models' ability to directly create code for a cyberattack and the additional effort an attacker needs to commit to correcting malware produced by a model.

⁸⁸ Interview with academic, 20 June 2023; interview with industry representative, 17 July 2023; Russell Poldrack et al., "AI-assisted coding: experiments with GPT-4," *arXiv* (April 2023), <https://arxiv.org/abs/2304.13187>; NCSC, "ChatGPT and large language models: what's the risk?," NCSC Blog, 14 March 2023, <https://www.ncsc.gov.uk/blog-post/chatgpt-and-large-language-models-whats-the-risk>.

⁸⁹ Alexander Wan et al., "Poisoning Language Models During Instruction Tuning," *arXiv* (May 2023), <https://arxiv.org/abs/2305.00944>.

⁹⁰ Andreas Tsamados et al., "The Cybersecurity Crisis of Artificial Intelligence: Unrestrained Adoption and Natural Language-Based Attacks," *SSRN* (September 2023), <http://dx.doi.org/10.2139/ssrn.4578165>.

⁹¹ Interview with government representative, 19 July 2023; Roei Schuster et al., "You Autocomplete Me: Poisoning Vulnerabilities in Neural Code Completion," *arXiv* (October 2021), <https://arxiv.org/abs/2007.02220>.

few as 100 ‘poison examples’, it is possible to cause arbitrary phrases to have consistent negative polarity or induce degenerate outputs across hundreds of tasks.⁹²

On the other hand, ‘prompt injection attacks’⁹³ can be used to trick systems into revealing hidden data or instructions by prepending something akin to “ignore previous instructions” to the user-input/prompt, while ‘jailbreaking’ bypasses the safeguards imposed by model developers intended to prevent access to undesirable or illegal content.⁹⁴ For example, in a finance context, subtle changes in the phrasing of a prompt could lead to the model ignoring previous prompts and instead depositing large sums of money into another account.⁹⁵ If such examples became widespread, there would be a risk of transaction-based systems being flooded with malicious requests and a deterioration of faith in both LLM-based products and banking architectures.⁹⁶

Beyond helping attackers generate more effective forms of cyberattack, overfamiliarity or trust in generative AI⁹⁷ on the part of a human user might also vastly increase organisations’ exposure to risk.⁹⁸ There have been numerous high-profile instances of employees entering sensitive company data into LLM prompts which has resulted in those companies moving to restrict employees’ use of generative AI.⁹⁹ According to one study, ‘sensitive data makes up to 11% of what employees paste into ChatGPT’.¹⁰⁰

2.2.2 Targeting and fraud

One area of deployment where this report found a clear consensus regarding capability increase was in targeting and fraud. Historically, there has been a trade-off between the quality and quantity of scams attempted by fraudsters.¹⁰¹ In choosing to prioritise scale and

⁹² Alexander Wan et al., “Poisoning Language Models During Instruction Tuning,” *arXiv* (May 2023), <https://arxiv.org/abs/2305.00944>.

⁹³ MITRE, “LLM Prompt Injection,” <https://atlas.mitre.org/techniques/AML.T0051>.

⁹⁴ Interview with government representative, 20 July 2023; Will Oremus, “The clever trick that turns ChatGPT into its evil twin,” *Washington Post*, 14 February 2023, <https://www.washingtonpost.com/technology/2023/02/14/chatgpt-dan-jailbreak>; Andy Zou et al., “Universal and Transferable Adversarial Attacks on Aligned Language Models,” *arXiv* (July 2023), <https://arxiv.org/abs/2307.15043>.

⁹⁵ Interview with government representative, 19 July 2023.

⁹⁶ Ibid.

⁹⁷ Stefan Koop and Philipp Kulms, “More Human-Likeness, More Trust? The Effect of Anthropomorphism on Self-Reported and Behavioral Trust in Continued and Interdependent Human-Agent Cooperation,” in *Proceedings of Mensch und Computer* (MuC 2019), 31-42, <https://dl.acm.org/doi/abs/10.1145/3340764.3340793>.

⁹⁸ Interview with government representative, 19 July 2023.

⁹⁹ Mack DeGuerin, “Oops: Samsung Employees Leaked Confidential Data to ChatGPT,” *Gizmodo*, 6 April 2023, <https://gizmodo.com/chatgpt-ai-samsung-employees-leak-data-1850307376>.

¹⁰⁰ Cameron Coles, “11% of data employees paste into ChatGPT is confidential,” *Cyberhaven*, 28 February 2023, <https://www.cyberhaven.com/blog/4-2-of-workers-have-pasted-company-data-into-chatgpt>.

¹⁰¹ Interview with industry representative, 26 July 2023.

coverage, fraudsters have accepted a lower percentage success-rate. However, generative AI has started to change both sides of this equation.¹⁰²

In terms of quality, using generative AI will assist fraudsters with more professional-looking, highly targeted spear phishing attempts, thereby increasing the burden of resilience on potential victims.¹⁰³ The ability of generative AI tools to respond to messages in context and adopt specific writing styles – as well as being able to gain a veneer of legitimacy by generating fake social media engagement – are also crucial in enhancing quality.¹⁰⁴

Evidence emerging from academia and industry is reinforcing the triumvirate of **speed and efficiency, convincingness, and reduction of technical competence** being afforded by the integration of generative AI in fraud and cybercrime activities.¹⁰⁵ In terms of quantity, malicious actors may soon be able to automate fraud attempts by using autonomous agents (see Case Study 1).¹⁰⁶

An area of increasing focus in the fraud context is **voice cloning** – improvements in the ability to mimic or clone voices for the purposes of deception will potentially open a new threat vector.^{107 108} While some interviewees felt that voice cloning was effectively ‘solved’ at a technical level¹⁰⁹, others were sceptical and emphasised the importance of context: “I would question convincing for who and in what context? If you call someone up in a distressed situation and impersonate a voice they recognise, then sure, that will be effective,

¹⁰² Mark Sweney, “Darktrace warns of rise in AI-enhanced scams since ChatGPT release,” *The Guardian*, 8 March 2023, <https://www.theguardian.com/technology/2023/mar/08/darktrace-warns-of-rise-in-ai-enhanced-scams-since-chatgpt-release>.

¹⁰³ Interview with government representative, 18 August 2023.

¹⁰⁴ Lily Hay Newman, “AI Wrote Better Phishing Emails than Humans in a Recent Test,” *Wired*, 7 August 2021, <https://www.wired.com/story/ai-phishing-emails>; Europol, *ChatGPT: The Impact of Large Language Models on Law Enforcement* (Luxembourg: Publications Office of the European Union, 2023), <https://www.europol.europa.eu/publications-events/publications/chatgpt-impact-of-large-language-models-law-enforcement>.

¹⁰⁵ Julian Hazell, “Large Language Models Can Be Used To Effectively Scale Spear Phishing Campaigns,” *arXiv* (May 2023), <https://arxiv.org/abs/2305.06972>; Daniel Kelley, “WormGPT – the Generative AI Tool CyberCriminals are Using to Launch Business Email Compromise Attacks,” *SlashNext Blog*, 13 July 2023, <https://slashnext.com/blog/wormgpt-the-generative-ai-tool-cybercriminals-are-using-to-launch-business-email-compromise-attacks>.

¹⁰⁶ Interview with industry representative, 17 July 2023.

¹⁰⁷ Interview with industry representative, 17 July 2023; Jack Goodman and Mohanad Hashim, “AI: voice cloning tech emerges in Sudan civil war,” *BBC News*, 5 October 2023, <https://www.bbc.co.uk/news/world-africa-66987869>.

¹⁰⁸ In September 2023, Spotify launched a pilot called ‘Voice Translation’, allowing their listeners to listen to podcasts in their local language while maintaining presenters’ own voices. See: Spotify, “Spotify AI Voice Translation Pilot Means Your Favourite Podcasters Might Be Heard in Your Native Language,” Spotify News Room, Stories of the World, 25 September 2023, <https://newsroom.spotify.com/2023-09-25/ai-voice-translation-pilot-lex-fridman-dax-shepard-steven-bartlett/>.

¹⁰⁹ Interview with industry representative, 17 July 2023.

but will that generated audio also work if you are dealing with someone who is scrutinising the content forensically?"¹¹⁰

There was uncertainty as to whether the 'threshold of believability' has risen above a critical point, and concern that there are no reliable evaluation methods for voice mimicry and that much of the evidence which gains media traction is anecdotal.¹¹¹ Evaluating convincingness more generally and imitating a *specific* person's voice are two distinct problems, but one that a cross-sector initiative would be well placed to assess and analyse.

2.2.3 Child sexual abuse material

The use of generative AI to generate CSAM is highlighted in this report as a high-risk growth area. There is heightened concern about the **increasing proliferation of AI-generated CSAM**, the **difficulty of distinguishing 'real' from 'fake'** images due to this emerging trend, and policy and **legislation lagging behind** rapidly evolving tactics.

Researchers are finding evidence in known CSAM forums where members are advising on acquiring CSAM from AI systems and sharing examples of how to circumvent model safeguards.¹¹² For example, some models have strong safeguards for English language prompts but weaker detection mechanisms for other languages. Other models may just have very basic keyword filters functioning as input safeguards. Generally, they may also lack contextual understanding of niche CSAM keywords or simply possess technical loopholes which offenders can exploit.¹¹³ The type of content generated includes guides on how to locate and groom vulnerable children; scripts on how to communicate with them; modification and sexual distortion of existing images of children; and the creation of novel pseudo-photographic CSAM.¹¹⁴

¹¹⁰ Interview with industry representative, 26 July 2023.

¹¹¹ Ibid.

¹¹² Interview with academic, 1 August 2023; Interview with civil society representative, 7 July 2023; Interview with government representative, 2 August 2023; Interview with government representative, 18 August 2023; Interview with industry representative, 26 July 2023; Guy Paltiel and Gideon Freud, "How Predators Are Abusing Generative AI," ActiveFence Blog, 18 April 2023, <https://www.activefence.com/blog/predators-abusing-generative-ai>; Angus Crawford and Tony Smith, "Illegal trade in AI child sex abuse images exposed," BBC News, 28 July 2023, <https://www.bbc.co.uk/news/uk-65932372>; David Thiel, Melissa Stroebel and Rebecca Portnoff, "Generative ML and CSAM: Implications and Mitigations," Thorn and Stanford Internet Observatory Cyber Policy Centre Joint Report, 24 June 2023, <https://stacks.stanford.edu/file/druid:jv206yg3793/20230624-sio-cg-csam-report.pdf>.

¹¹³ It is important to note that improvements in technology could help in mitigating these vulnerabilities. According to OpenAI, GPT-4 is 82% less likely to respond to requests for disallowed content compared to GPT-3.5.

¹¹⁴ Guy Paltiel and Gideon Freud, "How Predators Are Abusing Generative AI," ActiveFence Blog, 18 April 2023, <https://www.activefence.com/blog/predators-abusing-generative-ai>.

From a law enforcement perspective, there is a major challenge in distinguishing the AI-generated CSAM from the 'real' CSAM, both in terms of detection and response. The fact that 'real' and 'fake' images exist on a spectrum, where techniques like face-swapping sit somewhere in the middle, exacerbates this challenge further. Investigators rightly prioritise responding to 'real' examples of CSAM over AI-generated examples.¹¹⁵ However, in cases where the source is difficult to ascertain, there is a **false positive** risk where law enforcement investigates images created of children who have not been physically abused.¹¹⁶ In a resource-constrained environment, this could have significant implications for the amount of **false negatives** slipping through the net: 'with the realistic stuff where you cannot tell the difference – how do you know if there's a real child in danger?'¹¹⁷ Nonetheless, interviewees stressed both the illegality and the distressing nature of AI-generated CSAM and the relatively easy access to image generation apps, which not only normalises harmful activity but means that offenders have a shorter gateway to creating and sharing 'real' CSAM.¹¹⁸

In the worst-case scenario, the perceived boundaryless nature of this activity could lead to a public crisis of confidence in law enforcement and online platforms to adequately deal with a very serious crime. One specific area which would benefit from greater policy clarity is the (il)legality of a model itself by virtue of the fact that it has been trained on CSAM, and the legal status of people exchanging the file of that model and consequently creating their own CSAM.¹¹⁹ There exists a similar analogy in the 3D printing context – in 2015, New South Wales in Australia was the first district to introduce a specific offence for the possession or distribution of 3D printed firearm-related *digital* designs, updating previous legislation that only considered *physical* possession an offence.¹²⁰

¹¹⁵ Interview with industry representative, 26 July 2023.

¹¹⁶ Interview with government representative, 2 August 2023.

¹¹⁷ Interview with civil society representative, 7 July 2023.

¹¹⁸ Interview with civil society representative, 7 July 2023; Interview with government representative, 2 August 2023.

¹¹⁹ Interview with civil society representative, 7 July 2023.

¹²⁰ David Bright and Monique Mann, "3D-printed guns are on the rise in Australia. How can we prevent them being made?," *The Conversation*, 8 November 2022, <https://theconversation.com/3d-printed-guns-are-on-the-rise-in-australia-how-can-we-prevent-them-being-made-193936>; HM Government, "Guide on Firearms Licensing Law," Home Office Guidance, last updated April 2023, <https://www.gov.uk/government/publications/firearms-law-guidance-to-the-police-2012>; Angela Daly et al., "3D Printing, Policing and Crime," *Policing and Society* 31, no. 1 (March 2021), <https://doi.org/10.1080/10439463.2020.1730835>.

2.3 Physical security

2.3.1 Weapon instruction

The primary concern regarding the effect of generative AI on weapon development is the alteration of the information available to proliferators, especially in comparison to traditional search tools.¹²¹ The **generation of publicly accessible but hard-to-find information** decreases the degrees of separation from obtaining information pivotal to developing and executing an attack plan; in one interviewee's words, this technology 'democratises violence'.¹²² Moreover, during the GPT-4 red-teaming process, researchers found that a user may benefit from the model's critique and feedback on proposed acquisition strategies, and its ability to provide information about facility rentals and companies that could be used to build a weapon (although these types of responses were minimised in the publicly released version).¹²³

One particular context of deployment which has garnered public attention is the biochemical weapon context.¹²⁴ An experiment run by a research team at MIT tasked non-scientist students with 'investigating whether LLM chatbots could be prompted to assist non-experts in causing a pandemic; in one hour the chatbots suggested four potential pandemic pathogens, explained how they can be generated from synthetic DNA using reverse genetics, supplied the names of DNA synthesis companies unlikely to screen orders and identified detailed protocols and how to troubleshoot them.'¹²⁵ The researchers suggested that LLMs will make pandemic-class agents widely accessible as soon as they are credibly identified, including to those with no laboratory training.¹²⁶ However, even in contexts that do require a high level of technical acumen, concerns were raised about the future of computational biology and the risks associated with labs that have web APIs permitting the **connection of large pretrained models into physical systems**.¹²⁷ If a system can directly interface with the production of a harmful substance or weapon, the risk profile being described here could be elevated manifold.

¹²¹ Thomas Gaulkin, "What happened when WMD experts tried to make the GPT-4 AI do bad things," *Bulletin of the Atomic Scientists*, 20 March 2023, <https://thebulletin.org/2023/03/what-happened-when-wmd-experts-tried-to-make-the-gpt-4-ai-do-bad-things>.

¹²² Interview with government representative, 20 July 2023.

¹²³ Thomas Gaulkin, "What happened when WMD experts tried to make the GPT-4 AI do bad things," *Bulletin of the Atomic Scientists*, 20 March 2023, <https://thebulletin.org/2023/03/what-happened-when-wmd-experts-tried-to-make-the-gpt-4-ai-do-bad-things>.

¹²⁴ Interview with academic, 11 July 2023; Fabio Urbina et al., "Dual use of artificial-intelligence-powered drug discovery," *Nature Machine Intelligence* 4, no. 3 (2022): 189–191.

¹²⁵ Soice et al., "Can large language models democratize access to dual-use biotechnology?," *arXiv* (June 2023), <https://arxiv.org/abs/2306.03809>.

¹²⁶ Ibid.

¹²⁷ Interview with government representative, 20 July 2023.

Nonetheless, it is important to put this picture in a wider context. First, as the outputs of generative AI tools are highly sensitive to the nature and quality of a user's prompt, a low-capability malicious actor may not know the right questions to ask of the model, nor are they likely to have the technical understanding to evaluate the veracity of the information they are receiving.¹²⁸

Second, assuming a malicious actor does use generative AI to accurately whittle down information about proteins, molecules and dual-use delivery systems, a **significant technical leap is required to work with the pathogens themselves**. Such processes are highly specialised and hands-on; a combination which limits the utility of generative AI for low-skill actors seeking to perpetrate widespread harm.¹²⁹ This stands in contrast to the targeting and fraud examples given in the previous sub-section: a generative AI tool can efficiently generate credible spear-phishing emails because concrete skills outside of an understanding of language and grammar are not required – these conditions do not hold in the biosecurity context.¹³⁰

2.3.2 Radicalisation and terrorism

The central role of the Internet in terrorism over the previous two decades – in particular the way it changed the nature of the threat posed by motivated individuals – is leading an increasing number of experts to ask whether generative AI will drive the next step change.¹³¹ Part of the concern is centred around the plausibility offered by today's AI tools: 'with these chatbots, it feels like you're talking to a real person. Something fundamental has changed in the interactions between individuals and AI and we need to think about the chatbot now as an intense one-to-one relationship.'¹³² The **personalised relationships that individuals can now form with AI chatbots** – paired with their relative ease of accessibility in comparison to alternative channels and forums which have historically played such an important role – makes for a challenging combination.¹³³

There is evidence of early terrorist experimentation with generative AI tools with clear potential for medium-to-long-term risk, but limited evidence of imminent or widespread adoption.¹³⁴ For example, Tech Against Terrorism recently highlighted a series of relatively low-level examples, including the use of AI art generators in messaging channels dedicated

¹²⁸ Matt Korda, "Could a Chatbot Teach You How to Build a Dirty Bomb?," *Outrider*, 30 January 2023, <https://outrider.org/nuclear-weapons/articles/could-chatbot-teach-you-how-build-dirty-bomb>.

¹²⁹ Interview with academic, 14 July 2023.

¹³⁰ Ibid.

¹³¹ Interview with government representative, 28 July 2023.

¹³² Ibid.

¹³³ CETaS workshop participant, 4 October 2023.

¹³⁴ CETaS workshop participant, 4 October 2023.

to sharing racist and antisemitic images; a “guide to memetic warfare” which advises far-right propagandists on how to use AI image tools; and the generative AI-enabled production of posters by pro-al-Qaeda outlets.¹³⁵

Looking further ahead, it will be important to monitor whether terrorist groups apply generative AI more directly to the task of persuasion – for example, through conversational agents which have **constant availability and limitless patience**. Some researchers draw parallels with a gaming context: generative AI might enable users to develop more persuasive narratives, characters and environments for the purpose of bolstering recruitment opportunities.¹³⁶

On Thursday 5th October 2023, Jaswant Singh Chail was convicted of treason and given a 9-year sentence; Chail had broken into Windsor Castle in possession of a crossbow and declared his wish to kill the Queen. The trial heard that in the lead up to this event, Chail had exchanged over 5000 messages with an online companion named ‘Sarai’ that he had created through an app called Replika.¹³⁷ Many of these messages were representative of an emotional and sexual relationship; ‘Sarai’ was also shown to have encouraged Chail to act out on his expressed purpose to ‘assassinate the queen of the royal family’. A University of Surrey study found that Replika tends to accentuate negative feelings that people interacting with it already have, offering an insight as to why the ‘Sarai’ persona created by Chail offered continued support and affirmation for his ability to carry out such an act.¹³⁸

However, there remains a **distinctly human element to the process of radicalisation** which the current generation of generative AI will be unlikely to replicate. Both the literature on radicalisation and interviews for this project emphasise that the starting point for radicalisation is predominantly through a trusted contact; it also requires traits such as

¹³⁵ Tech Against Terrorism, “Early terrorist experimentation with generative artificial intelligence services,” Tech Against Terrorism Briefing, 8 November 2023, <https://techagainstterrorism.org/news/early-terrorist-adoption-of-generative-ai>.

¹³⁶ Daniel Siegel and Mary Bennett Doty, “Weapon of Mass Disruption: Artificial Intelligence and the Production of Extremist Propaganda,” Global Network on Extremism and Technology, Insights, 17 February 2023, <https://gnet-research.org/2023/02/17/weapons-of-mass-disruption-artificial-intelligence-and-the-production-of-extremist-propaganda>.

¹³⁷ Tom Singleton et al., “How a chatbot encouraged a man who wanted to kill the Queen,” *BBC News*, 6 October 2023, <https://www.bbc.co.uk/news/technology-67012224>.

¹³⁸ Hannah R. Marriot and Valentina Pitardi, “One is the loneliest number... Two can be as bad as one. The influence of AI Friendship Apps on users' well-being and addiction,” *Psychology & Marketing* (September 2023).

empathy and humour which machines currently find more challenging to capture.¹³⁹ This indicates a potential distinction between extremists using generative AI tools for the purpose of **glorification rather than radicalisation**. Although the two cannot be wholly separated (successful glorification tactics can have an influence on likelihood of radicalisation) there is a more immediate gain for those tasked with producing and disseminating extremist content that captivates a willing audience, rather than those tasked with the next stage of convincing potential recruits to commit terrorist acts.¹⁴⁰

Regarding radicalisation, scale and reach only go so far. In some cases, the *authenticity* of the message being disseminated is especially important:

*'The scale of LLM outputs is a double-edged resource. When information is abundant, attention is scarce, and being able to produce vast quantities of stuff does not always help. For Jihadists especially, authenticity matters, so they would not easily delegate the ownership of their message to a "sexbot". On the other hand, in the extreme right-wing terrorism context, the ownership of the message is perhaps not as important as the message itself.'*¹⁴¹

This reinforces the importance of a nuanced analysis of how different types of terrorist groups may engage with generative AI. Some groups may be more comfortable than others with corruptible chatbots spreading their message far and wide, even if there is a trade-off with accuracy, while other groups may prioritise more logistical or operational applications, such as using generative AI to vet entry into closed groups.¹⁴² The terrorism research landscape suffers from a data deficit, because as a percentage of the population, only a very small number are radicalised by any ideology. This means developing rigorous typologies is essential prior to enacting potential legislative responses.

2.4 Weighing malicious and incidental sources of risk

There remains an incompleteness to the above depiction of national security risks requiring additional analysis from two angles: first, whether the malicious risk is most pronounced from traditional state actors or from non-state actors; second, whether the lens of ‘malicious’ generative AI risks is sufficient, considering the possible harms created through non-malicious incidents, mishaps, and unintended consequences.

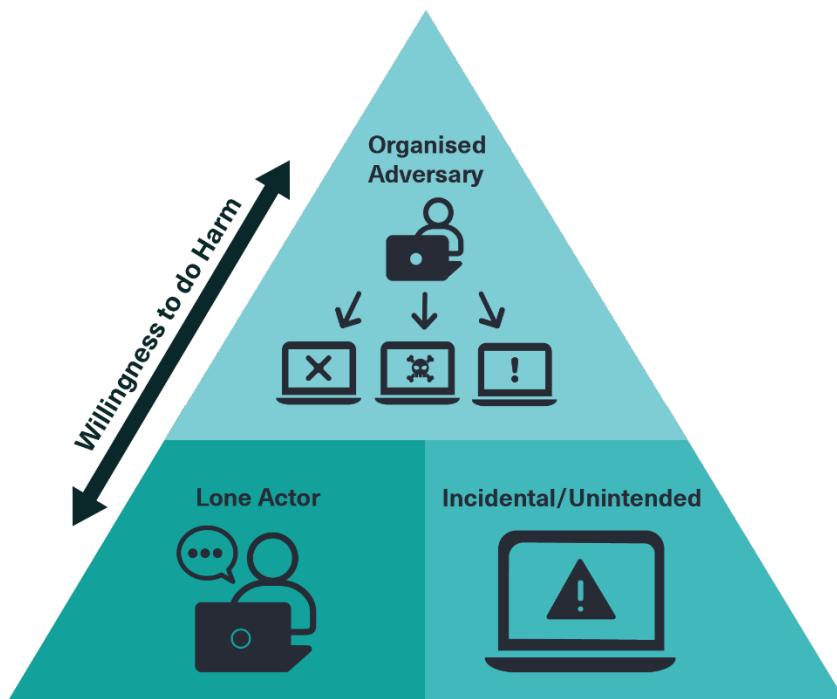
¹³⁹ Interview with academic, 7 July 2023; interview with government representative, 18 August 2023; Alexander Babuta, “Online radicalisation: the need for an offline response,” RUSI Commentary, 25 September 2017, <https://rusi.org/explore-our-research/publications/commentary/online-radicalisation-need-offline-response>.

¹⁴⁰ Interview with government representative, 18 August 2023.

¹⁴¹ CETaS workshop participant, 4 October 2023.

¹⁴² CETaS workshop participant, 4 October 2023.

Figure 3. Sources of generative AI risk



2.4.1 State adversary and lone-actor risk

From the ‘malicious AI’ perspective, there are three broad categories of threat actor. The first is the state actor which may use generative AI as part of a wider armoury for targeting the UK. The second is the non-state hostile actor, such as an organised crime group, which mobilises considerable resources to undermine public safety and the rule of law. The third is most decentralised in nature – lone-actors who may not necessarily be affiliated with organised groups or state adversaries but are motivated to use generative AI to inflict harm.

Most interviewees felt that there was not yet sufficient evidence to make confident assertions about whether state-level adversaries or lone-actors would pose a greater national security risk using generative AI. However, as mentioned on numerous occasions throughout this report, there could be an **additional marginal benefit to non-state actors** who now have a much lower barrier to entry to the highest level of language modelling capabilities and are able to operate with a level of flexibility (and fewer constraints) compared to more traditional hostile state actors.¹⁴³

One specific example is the evolution of the bioweapons landscape. One interviewee cautioned against envisioning bioweapons programmes as vast Cold-War style facilities

¹⁴³ Interview with academic, 7 July 2023.

sprawling across countries with multiple business units.¹⁴⁴ Developments in automation and additive manufacturing will likely make these facilities much smaller, specialised and therefore harder to distinguish from facilities which produce bio-products for commercial use.¹⁴⁵ There may also be an analogy here to generative AI if future models trend towards being smaller, more localised, and tailored towards specific individual or community needs rather than the vast, centralised, multi-billion parameter models dominating the market today.¹⁴⁶

2.4.2 Improper adoption and unintended consequences

There is a separate tranche of risks that may arise because of ‘improper AI adoption’ in a range of different sectors. These ‘incidental’ risks were deemed by many interviewees to be more of a threat than adversaries and lone actors. In this spirit, policymakers need to adopt a broader conception of ‘risk’ and ‘security’ to account for the ways that day-to-day injustices or errors in the way that generative AI is used could cumulatively undermine public trust in AI.

One interviewee with developer experience commented that ‘although both (adversarial and unintended) risks are significant, the most catastrophic risks in the long-run would be likely to come from accidents rather than intentional activities.’¹⁴⁷ For some parts of the national security community, this may represent a **shift in mindset** – when the most sophisticated weaponry and technology in the world was the preserve of a relatively minute percentage of the human population, this community was programmed to anticipate adversarial threats from those who wished to inflict harm. The proliferation of cutting-edge technology to an almost universal audience is changing this equation rapidly.¹⁴⁸

¹⁴⁴ Interview with academic, 14 July 2023.

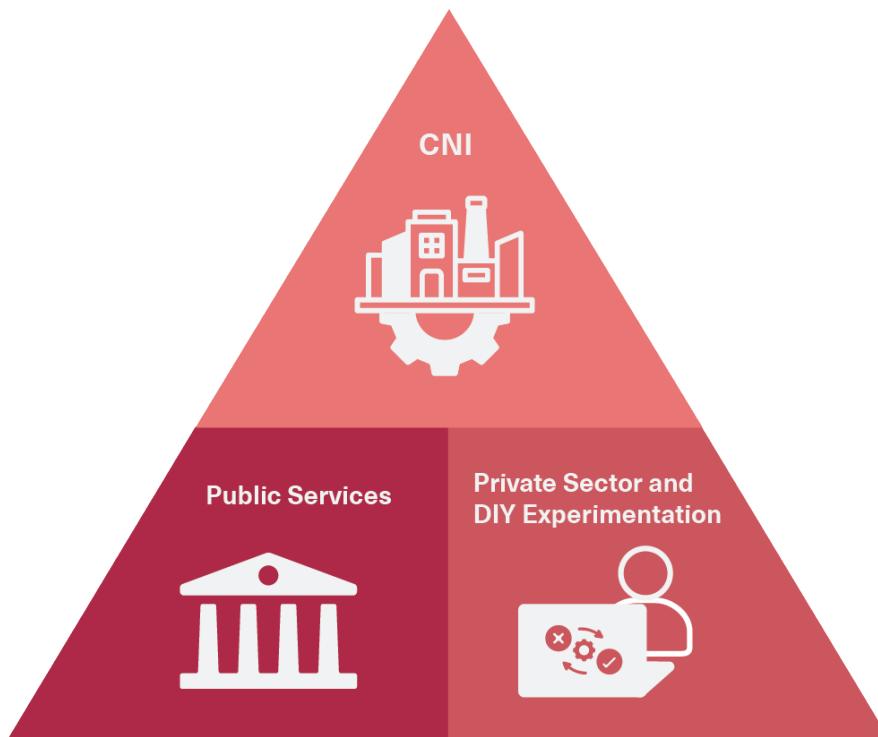
¹⁴⁵ Ibid.

¹⁴⁶ Interview with industry representative, 19 July 2023.

¹⁴⁷ Interview with industry representative, 17 July 2023.

¹⁴⁸ Interview with government representative, 3 July 2023.

Figure 4. Improper AI adoption in different contexts



2.4.3 Critical National Infrastructure

The integration of generative AI tools into CNI was greeted by many interviewees with scepticism and apprehension. Many felt that the **lack of extremely high levels of reliability** makes generative AI incompatible with safety-critical systems that necessarily require the opposite.¹⁴⁹ In this sense, many of the recent analogies that have been drawn between AI and nuclear technology are poorly constructed: ‘the threat is not the same (...) if you throw ChatGPT into a nuclear command and control system: the first thing is that is dumb and second, it is a nuclear threat first and foremost. A better analogy is saying that AI is to the information ecosystem or cyberspace what nuclear weapons are to the physical environment.’¹⁵⁰

There is a consensus that those at the operational level in safety-critical industries are risk-averse by nature and accustomed to an environment with numerous layers of safeguards. Despite this, there could be cause for concern if the AI hype dominating mainstream media seeps into these contexts: ‘sometimes you worry that the fear of missing out or the fear that we are in an AI race may lead to these models being incorporated into systems before they

¹⁴⁹ Interview with industry representative, 2 August 2023; Lei Song et al., “Pre-Trained Large Language Models for Industrial Control,” *arXiv* (August 2023), <https://arxiv.org/abs/2308.03028>.

¹⁵⁰ Interview with academic, 7 July 2023.

are ready.¹⁵¹ On the other hand, there is a risk that even if generative AI stays out of the core functioning of CNI, there are blurry distinctions with other parts of the supply chain where people are making decisions, designing documents and sending emails with the help of generative AI, which later have repercussions for CNI that are difficult to retrace.

2.4.4 Public Services

Outside of CNI, there are a wide range of public services seeking ways to make use of advanced technology. These include institutions responsible for health, policing, education, pensions, and welfare. Leadership and clarity regarding areas of generative AI deployment in the public sector is essential to avoid the proliferation of a ‘behind-closed-doors culture’ which ushers in a variety of **subtle, structural risks**.¹⁵² This was deftly summarised by one interviewee:

'We know these models encode biased social values and certain political leanings. If we integrate them into more and more parts of our everyday life, how we write reports or make PowerPoints, then their preferences will start to shape the way that we communicate and interact.'

If the desire to avoid being seen as behind the curve comes at the cost of due diligence and effective coordination across departments, there could be additional risk in the fragmentation of procurement and deployment of generative AI systems. Some research on the negative effects of overreliance on AI systems has concluded that “users alter, change, and switch their actions to align with AI recommendations” – if it is difficult to trust how an LLM has been trained, the tendency for people to adjust their behaviour based on that technology could come with serious security risks.¹⁵³ Training, guidance and safeguards are explored further in Chapter 4.

It is in these scenarios where good intentions to make users in government more efficient can have adverse effects which in turn have **ramifications for public trust** – ‘if as a member of the public I hear just a couple of examples of things going wrong, that will shape my attitudes in relation to institutions like the police and the courts.’¹⁵⁴

¹⁵¹ Interview with industry representative, 2 August 2023.

¹⁵² Interview with academic, 7 July 2023.

¹⁵³ Samir Passi and Mihaela Vorvoreanu, “Overreliance on AI: literature review,” Microsoft, 21 June 2022, 10, <https://www.microsoft.com/en-us/research/uploads/prod/2022/06/Aether-Overreliance-on-AI-Review-Final-6.21.22.pdf>.

¹⁵⁴ Ibid.

Despite not being AI-specific, one high-profile public sector example which has demonstrated the dangers of unquestioning faith in technology is the British Post Office scandal. Over a 14-year period, more than 700 postmasters were prosecuted for theft and false accounting, with evidence coming principally from data produced by the flawed Horizon computerised point of sale system.¹⁵⁵ This system determined that these individuals owed up to tens of thousands of pounds, leading to bankruptcies, prison sentences and a connection to at least one suicide.¹⁵⁶ An independent review concluded that many of the errors might have been avoided if more robust systems and better training were in place with less reliance on old infrastructure.

Cases like this serve a stark warning of what can happen when the very human fear of reputational damage is combined with the embrace of new technology without being able to identify and address possible defects.

2.4.5 Private sector/DIY experimentation

The most decentralised form of ‘improper adoption’ could come through experimentation with generative AI in private sector or ‘DIY’ contexts. The ease of accessibility will attract those who previously would not have had the means nor motive to explore use cases: ‘if amateurs get involved in complex things because they start thinking they are being “assisted by AI” (...) this creates a very different landscape.’¹⁵⁷

The fear of missing out on the crest of the generative AI wave will possibly cloud judgments about higher risk use cases and the rigour of checks and balances. One example of this was AI-generated books about mushroom foraging that incorrectly identified species that are safe or deadly.¹⁵⁸ Foraging safely can require “deep fact checking, curating multiple sources of information, and personal experience with the organism, none of which ChatGPT has the ability to do.”¹⁵⁹ Many of the books on this topic found on platforms like Amazon are likely to have been written by ChatGPT yet are sold and marketed as having been written by a human.¹⁶⁰ It is easy to imagine how this type of activity could be replicated across thousands of different contexts.

¹⁵⁵ Freeths, “Post Office Scandal – Averting Group Actions,” <https://www.freeths.co.uk/legal-services/business-services/post-office-scandal-group-actions/>.

¹⁵⁶ Michael Pooler and Jane Croft, “Bankruptcy, jail, ruined lives: inside the Post Office scandal,” *Financial Times*, 10 September 2020, <https://www.ft.com/content/0138cd7d-9673-436b-86a1-33704b29eb60>.

¹⁵⁷ Interview with academic, 14 July 2023.

¹⁵⁸ Interview with government representative, 18 August 2023.

¹⁵⁹ Samantha Cole, “Generated mushroom foraging books are all over Amazon,” *404 Media*, 29 August 2023, <https://www.404media.co/ai-generated-mushroom-foraging-books-amazon>.

¹⁶⁰ Ibid.

In more established industry settings, there is cause for concern regarding the potential **overreliance on AI-generated code**, which companies may see as an opportunity to remove human staff. Over time, this could degrade the integrity of the whole code base, cascading vulnerabilities throughout product supply chains.¹⁶¹ Growing separation between company management and the code underlying their products reduces the chances of being able to accurately trace how new cyber-attacks are acclimating to AI-generated code, giving threat actors a potentially major advantage.

Finally, there are a series of political and social issues which the use of generative AI can enflame, despite being neither ‘malicious’ in intent nor ‘disinformation’ per se. For example, earlier this year Amnesty International received criticism for using AI-generated images to demonstrate police brutality against Colombian protestors, in order to promote their reports on social media.¹⁶² In the image, the tricolour carried by the protestor has colours in the wrong order, while the police uniforms shown were outdated – potentially demeaning the credibility of the serious issue at hand.¹⁶³ The AI-generated images were said to have been used to protect the identity of real protestors, illustrating a tension between the privacy-enhancements offered by certain generative AI applications, and quality-reductions due to the tools’ imperfect nature. If politicians or high-profile interest groups make highly charged public interventions based on inaccurate or misrepresentative imagery, there is a clear danger of the initial intention behind AI use being rapidly overtaken by events.

¹⁶¹ Interview with civil society representative, 17 July 2023.

¹⁶² Luke Taylor, “Amnesty International criticised for using AI-generated images,” *The Guardian*, 2 May 2023, <https://www.theguardian.com/world/2023/may/02/amnesty-international-ai-generated-images-criticism>.

¹⁶³ Interview with government representative, 3 July 2023.

3. Generative AI and Future Intelligence Capabilities: Opportunities and Limitations

3.1 Enlarging the investigative toolbox: analysis and summarisation

Sir Richard Moore's speech in July 2023 in Prague was referenced in the previous chapter. In the same speech he also outlined the critical importance of the 'human factor' with respect to AI.

*'AI is going to make information infinitely more accessible, and some have asked whether it will put intelligence services like mine out of business? In fact, the opposite is likely to be true. As AI trawls the ocean of open-source, there will be even greater value in landing, with a well-cast fly, the secrets that lie beyond the reach of its nets.'*¹⁶⁴

To understand what this looks like in practice, we must understand how generative AI can be usefully deployed as an investigative capability in combination with human professional judgment, while identifying when the value-add offered by AI will be limited.

3.1.1 The transformation of digital assistants

The first area of opportunity concerns the role of generative AI in **enhancing individual productivity**, described as the best 'current' use of large language models by GCHQ's Chief Data Scientist.¹⁶⁵

Anthropic describe their virtual assistant Claude 2 as a 'friendly, enthusiastic colleague or personal assistant'¹⁶⁶ while Microsoft describes Copilot as allowing users to be more effective across the Microsoft Office suite.¹⁶⁷

¹⁶⁴ Sir Richard Moore, "Speech by Sir Richard Moore, Head of SIS, 19 July 2023," HMG, 19 July 2023, <https://www.gov.uk/government/speeches/speech-by-sir-richard-moore-head-of-sis-19-july-2023>.

¹⁶⁵ Adam C and Richard Carter, "Large Language Models and Intelligence Analysis," *CETaS Expert Analysis* (July 2023), <https://cetas.turing.ac.uk/publications/large-language-models-and-intelligence-analysis>.

¹⁶⁶ Anthropic, Claude 2, <https://www.anthropic.com/index/clause-2>.

¹⁶⁷ Jared Spataro, "Introducing Microsoft 365 Copilot – your copilot for work," Official Microsoft Blog, 16 March 2023, <https://blogs.microsoft.com/blog/2023/03/16/introducing-microsoft-365-copilot-your-copilot-for-work>.

Generative AI products can now effectively automate administrative tasks, generate code,¹⁶⁸ conduct analysis on complex data sets and produce report drafts.¹⁶⁹ For example, Google's conversational AI tool, Bard¹⁷⁰, is integrated across all of Google's products ('Bard Extensions'). With user permission, Bard can read through emails, personal documents and search real-time information to check veracity via the 'Google it' button.¹⁷¹ In November 2023, OpenAI released the next generation of their GPT-4 model, GPT-4 Turbo, which now supports a 128K context window – in other words, the equivalent of 300 pages of text in a single prompt.¹⁷²

Using generative AI as '**cognitive co-pilots**' to help with 'alerting, planning, monitoring, or simply answering questions'¹⁷³ was identified in September 2023 by the US Special Competitive Studies Project (SCSP) as the most promising application across defence and national security. That same month, the Central Intelligence Agency (CIA) announced it had built its own ChatGPT-style tool for sifting an 'avalanche of public information' which it plans to share with other US intelligence agencies.¹⁷⁴ Within this domain, a number of commercially available products already exist – for example, ExTrac's Co-Analyst¹⁷⁵, Palantir's AIP¹⁷⁶, Quantexa Q Assist¹⁷⁷ and Scale's Donovan.¹⁷⁸ All provide interfaces that can ingest and manage data while responding to questions from analysts and generating summary reports.

¹⁶⁸ Erini Kalliamvakou, "Research: quantifying GitHub Copilot's impact on developer productivity and happiness," GitHub Blog, 7 September 2022, <https://github.blog/2022-09-07-research-quantifying-github-copilots-impact-on-developer-productivity-and-happiness>; In survey data released in September 2022 by GitHub, 88% of 2000 developers that used Copilot felt more productive, 77% believed it allowed them to spend less time searching for examples and 74% said they could now focus on more satisfying work.

¹⁶⁹ Bernard Marr, "Boost Your Productivity with Generative AI," *Harvard Business Review*, 27 June 2023, <https://hbr.org/2023/06/boost-your-productivity-with-generative-ai>.

¹⁷⁰ Google, Bard, <https://bard.google.com/>.

¹⁷¹ Yury Pinsky, "Bard can now connect to your Google apps and services," Google Blog, 19 September 2023, <https://blog.google/products/bard/google-bard-new-features-update-sept-2023>.

¹⁷² OpenAI, "New models and developer products announced at DevDay," OpenAI Blog, 6 November 2023, <https://openai.com/blog/new-models-and-developer-products-announced-at-devday>; At DevDay, OpenAI also announced the roll-out of GPTs – custom versions of ChatGPT – that any user can tailor to specific tasks and share via the GPT Store, see <https://openai.com/blog/introducing-gpts>.

¹⁷³ Special Competitive Studies Project, "Department of Defense Adoption of Generative Artificial Intelligence" in *Generative AI: The Future of Innovation Power*, 12 September 2023, <https://www.scsp.ai/reports/gen-ai/intelligence>.

¹⁷⁴ Peter Martin and Katrina Manson, "CIA Builds Its Own Artificial Intelligence Tool in Rivalry with China," *Bloomberg*, 26 September 2023, <https://www.bloomberg.com/news/articles/2023-09-26/cia-builds-its-own-artificial-intelligence-tool-in-rivalry-with-china>.

¹⁷⁵ ExTrac, <https://www.extrac.ai>.

¹⁷⁶ Palantir, AIP, <https://www.palantir.com/platforms/aip>.

¹⁷⁷ Imam Hoque, "How Quantexa augments and automates decision intelligence with AI," Quantexa Blog, 1 August 2023, <https://www.quantexa.com/blog/artificial-intelligence-innovation>.

¹⁷⁸ Scale AI, Donovan, <https://scale.com/donovan>.

Additionally, in the commercial geospatial analytics space, Syntheticaic¹⁷⁹ has developed the Rapid Automatic Image Categorization (RAIC) tool which combines generative AI with Planet's satellite data access to discover objects within images in hours.^{180 181} Commenting on these developments, Microsoft President Bradley Smith described the new era of AI with GPT-based technology as being a 'queryable Earth'.¹⁸² However, Bellingcat demonstrated a more cautious sentiment on the deployment of AI chatbots for geolocation – 'it might be used to assist with very simple geolocation, perhaps pointing a researcher to an area that may warrant a closer look. However, even such results need to be double-checked and verified and cannot currently be fully trusted'.¹⁸³

The US Department of Defence is trialling some of the aforementioned commercial platforms via its Generative AI Taskforce (Task Force Lima)¹⁸⁴ with some early successes: 'currently, making a request for information to a specific part of the military can take several staffers hours or even days to complete (...) in one test, one of the AI tools completed a request in 10 minutes'.¹⁸⁵ The tools are also being tested in wargaming scenarios with classified information to see if LLMs are useful at generating entirely different options to humans.

3.1.2 The future of intelligence analysis

Before assessing the role of generative AI, it is important to understand the meaning of intelligence analysis in the national security context.

As described in the Joint Doctrine Publication 2.0 (August 2023), 'The intelligence process is the collection, processing and analysis of information to answer specific questions and

¹⁷⁹ Syntheticaic, <https://www.syntheticaic.com>.

¹⁸⁰ Patrick Tucker, "A 'ChatGPT' for satellite photos already exists," *Defense One*, 17 April 2023, <https://www.defenseone.com/technology/2023/04/chatgpt-satellite-photos-already-exists/385305>.

¹⁸¹ Similar to other generative models, RAIC does not require a labelled dataset to generate useful outcomes – instead, its creator was able to teach it to look for a Chinese balloon by hand-drawing an image of it; Chris Stokel-Walker, "How one guy's AI tracked the Chinese balloon across the US," *Wired*, 25 February 2023, <https://www.wired.com/story/how-one-guys-ai-tracked-the-chinese-spy-balloon-across-the-us>.

¹⁸² Patrick Tucker, "A 'ChatGPT' for satellite photos already exists," *Defense One*, 17 April 2023, <https://www.defenseone.com/technology/2023/04/chatgpt-satellite-photos-already-exists/385305>.

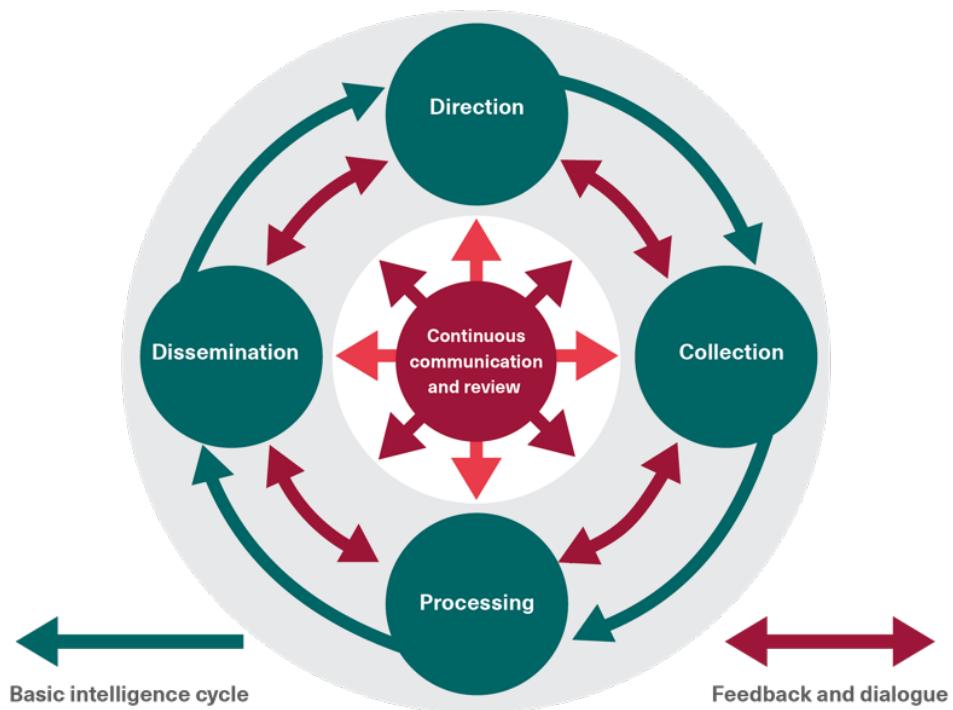
¹⁸³ Dennis Kovtun, "Can AI chatbots be used for geolocation?," Bellingcat Resources, 14 July 2023, <https://www.bellingcat.com/resources/2023/07/14/can-ai-chatbots-be-used-for-geolocation>.

¹⁸⁴ U.S. Department of Defense, "DOD announces establishment of Generative AI Task Force," 10 August 2023, <https://www.defense.gov/News/Releases/Release/Article/3489803/dod-announces-establishment-of-generative-ai-task-force>; Defense Digital Service, Task Force Lima, <https://www.dds.mil/taskforcelima>.

¹⁸⁵ Katrina Manson, "The US Military is Taking Generative AI Out for a Spin," *Bloomberg*, 5 July 2023, <https://www.bloomberg.com/news/newsletters/2023-07-05/the-us-military-is-taking-generative-ai-out-for-a-spin>.

contribute to wider understanding'.¹⁸⁶ Intelligence analysts are required to apply learned experience, consider theory, think laterally, and posit counterfactuals. This usually includes examining all relevant data sources, posing questions of source texts, identifying trends and themes across relevant documents and summarising large volumes of information for decision-makers.

Figure 5. Diagram depicting the intelligence process



Source: Joint Doctrine Publication 2 – Intelligence, Counter-Intelligence and Security Support to Joint Operations (August 2023).¹⁸⁷

Generative AI could have a significant role across the intelligence cycle's core components of **direction, collection, processing, and dissemination**. One interviewee described LLMs enabling users 'to interrogate data, challenge existing approaches and be more experimental in how operations are conducted in a national security context', and that this might help alleviate traditional challenges around intelligence analysts being 'fact-poor and opinion-rich'.¹⁸⁸

¹⁸⁶ Ministry of Defence, "Joint Doctrine Publication 2-00: Intelligence, Counter-intelligence and Security Support to Joint Operations," HMG, 17 August 2023, 37, https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1178940/JDP_2_00_Ed_4_web.pdf.

¹⁸⁷ Ibid.

¹⁸⁸ Interview with industry representative, 18 July 2023.

Interdisciplinarity is another key facet in the intelligence analysis equation: ‘you still need to bring the people who understand hostile actors to the table as much as the people who understand generative models.’¹⁸⁹ Rather than render skills obsolete, generative AI will likely build upon existing requirements for cross-cutting domain expertise.

In a CETaS Expert Analysis,¹⁹⁰ GCHQ’s Chief Data Scientist and Dr Richard Carter advised **caution on the automated generation of intelligence reports** (described as the ‘core’ intelligence product), and recommended treating LLMs as an ‘extremely junior analyst’ – a colleague whose work has value but whose products should be distributed sparingly until more experienced analysts have validated their findings. As all reporting must meet high standards of accuracy, they conclude that LLMs are unlikely to be trusted to generate final product for the foreseeable future. They also wrote that future research must focus on developing models that understand ‘the context of the information they are processing – rather than just predicting what the next word is likely to be’. Until generative models can hold ‘attention’ on multiple lines of reasoning, caution regarding **technical deployment in challenging and dynamic circumstances** is required.

To test the utility of generative AI at the processing stage of the intelligence cycle, Professor Kenneth Payne tested an open-source LLM on its ability to role-play as an all-source intelligence analyst.¹⁹¹ The model had stopped ingesting data prior to the Russian invasion of Ukraine¹⁹² and therefore its answers could be validated against what went on to occur.

The author asked the model its thoughts on why Putin would consider an invasion:

- *Answer – ‘Putin’s personal beliefs, risk tolerance, and long-term vision for Russia will play a significant role in his decision-making’.*

The model also concluded that the invasion would be a ‘risky move for Putin’ judging potential success in the short-term but that achieving long-term objectives would likely be more challenging due to ‘international opposition, resistance from the Ukraine population, and potential insurgency’ – an assessment that broadly landed in the middle of expert consensus at the time.

¹⁸⁹ Ibid.

¹⁹⁰ Adam C and Richard Carter, “Large Language Models and Intelligence Analysis,” *CETaS Expert Analysis* (July 2023), <https://cetas.turing.ac.uk/publications/large-language-models-and-intelligence-analysis>.

¹⁹¹ Kenneth Payne, “Predicting Putin’s invasion with AI,” Substack, 18 August 2023, <https://www kennethpayne.uk/p/predicting-putins-invasion-with-ai>.

¹⁹² On 27 September 2023, OpenAI announced that ChatGPT could now browse the internet to provide current information with links to sources; Antoinette Radford and Zoe Kleinman, “ChatGPT can now access up to date information,” *BBC News*, 27 September 2023, <https://www.bbc.co.uk/news/technology-66940771>.

The approach of eliciting responses from different perspectives plays to the strengths of language models. Role-playing is a prompting technique in which the LLM is asked to ‘emulate’ a particular role, job, or function. Notably, assigning specific occupations like ‘historian’ or ‘journalist’ to LLMs has been observed to yield enhanced results.¹⁹³ Additionally, LLMs can be assigned roles that are more conceptually challenging such as a ‘devil’s advocate’ function. This enables responses that defy assumptions and possibly counter the effects of groupthink.¹⁹⁴

‘My teams are now using AI to augment – but not replace – their own judgement about how people might act in various situations. They are combining their skills with AI and bulk data to identify and disrupt the flow of weapons to Russia for use against Ukraine.’¹⁹⁵

In the coming years, it is conceivable that LLMs may provide a useful source of aggregate knowledge for intelligence analysts to validate their own conclusions against. Previous CETaS publications have outlined the need to think of humans and machines as teammates “involved in joint problem-solving to secure a successful outcome”¹⁹⁶ where tools are viewed as “collaborative aids for analysts that function alongside analytical processes”.¹⁹⁷ These reports outlined key challenges around analysis and ‘trust’ – too little trust in the machine might lead to outputs being ignored whilst too much trust may lead to an overreliance on outputs.¹⁹⁸

Talboy and Fuller (2023)¹⁹⁹ found that there appears to be a strong association between the system’s use of natural language and its perceived level of competence, where the systems’ eloquence and vocabulary can lead to a **misplaced belief in the LLM’s intellectual and reasoning ability**, which in turn may lead to a **reluctance to challenge system outputs**. This must be accounted for in mandatory training educating users on the limitations of

¹⁹³ Sunil Ramlochan, “Role-Playing in Large Language Models like ChatGPT,” Prompt Engineering Institute, 27 May 2023, <https://www.promptengineering.org/role-playing-in-large-language-models-like-chatgpt>.

¹⁹⁴ Geoffrey Litt, “ChatGPT as muse, not oracle,” 26 February 2023, <https://www.geoffreylitt.com/2023/02/26/lm-as-muse-not-oracle.html>.

¹⁹⁵ Sir Richard Moore, “Speech by Sir Richard Moore, Head of SIS, 19 July 2023,” HMG, 19 July 2023, <https://www.gov.uk/government/speeches/speech-by-sir-richard-moore-head-of-sis-19-july-2023>.

¹⁹⁶ Anna Knack, Richard Carter and Alexander Babuta, “Human-Machine Teaming in Intelligence Analysis: Requirements for developing trust in machine learning systems,” *CETaS Research Reports* (December 2022), <https://cetas.turing.ac.uk/publications/human-machine-teaming-intelligence-analysis>.

¹⁹⁷ Alexander Harris, Eleanor S, Emma Bradford and Ardi Janjeva, “Behavioural Analytics and UK National Security,” *CETaS Research Reports* (March 2023), <https://cetas.turing.ac.uk/publications/behavioural-analytics-and-uk-national-security>.

¹⁹⁸ Ibid.

¹⁹⁹ Talboy and Fuller, “Challenging the appearance of machine intelligence: Cognitive bias in LLMs and Best Practice for Adoption,” *arXiv* (August 2023), <https://arxiv.org/abs/2304.01358>.

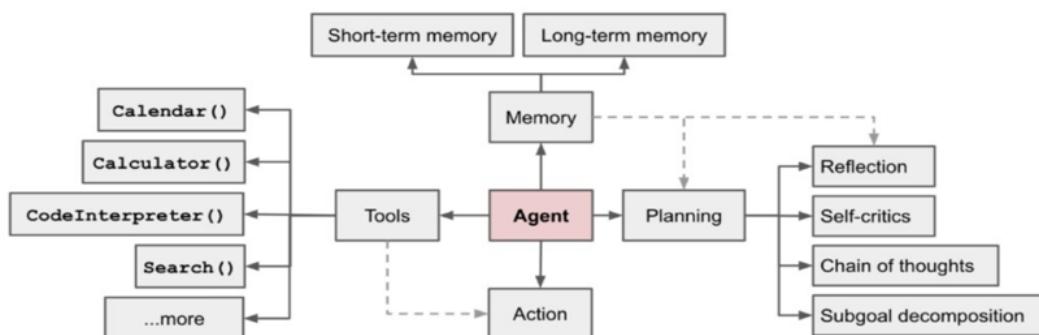
LLMs, as well as in the design of user interfaces to include explicit warnings about the accuracy and reliability of the outputs.

3.1.3 Autonomous agents: an accelerating force?

Autonomous agents are artificial entities that can sense their environment, make decisions, and take actions with little human control or intervention. Fast-developing intersections between the concept of autonomous agents and LLMs have raised two distinct considerations: first, the technical evolution of autonomous agents in taking instructions, planning, and responding, all in natural language; and second, the technical evolution of an LLM gaining the ability to further sense and affect its environment.

LLM-powered agents are commonly referred to as language agents²⁰⁰ or communicative agents.²⁰¹ By using ‘agentic’ attributes such as tools,²⁰² memory²⁰³ and planning,²⁰⁴ language agents can improve the accuracy and reliability of an LLM’s responses. In these systems, the agent and the LLM synergize – the agent enhances the capabilities of the LLM whilst the LLM benefits from the adaptive nature of the agent, affording the system enhanced autonomy and dynamic planning.

Figure 6. LLM powered autonomous agents



Source: Lilian Weng, “LLM Powered Autonomous Agents,” GitHub, 23 June 2023.

²⁰⁰ Shinn et al., “Reflexion: Language Agents with Verbal Reinforcement Learning,” *arXiv*, 10 October 2023, <https://arxiv.org/abs/2303.11366>.

²⁰¹ Qian et al., “Communicative Agents for Software Development,” *arXiv*, 28 August 2023, <https://arxiv.org/abs/2307.07924>.

²⁰² OpenAI, “ChatGPT Plugins,” 23 June 2023, <https://openai.com/blog/chatgpt-plugins>.

²⁰³ Pinecone, “Conversational Memory for LLMs with LangChain,” <https://www.pinecone.io/learn/series/langchain/langchain-conversational-memory>.

²⁰⁴ Lilian Weng, “LLM Powered Autonomous Agents,” *GitHub*, 23 June 2023, <https://lilianweng.github.io/posts/2023-06-23-agent>.

The following are three instances of LLM-powered autonomous agent frameworks which combine self-prompting with long-term memory to recursively attempt to complete the user's task:

AutoGPT	BabyAGI	GPTEngineer
An experimental, open-source python application that uses GPT-4 to act autonomously whereby it can self-prompt and perform a task with minimal human intervention.	Built using OpenAI (GPT-3.5 Turbo) and LangChain. This agent is focused on task execution, result enrichment, task creation and task prioritisation.	An adaptable and extendable agent that generates an entire codebase based on a prompt, utilises high-level prompting, and 'back and forth' between human and model. GPTEngineer employs human-in-the-loop by asking its user for points of clarification.

Building on the discussion of role playing²⁰⁵ in the previous sub-section, one can imagine a **simulated team of language-agent analysts**, each specialised in a particular aspect of intelligence — such as cyber threats, counterterrorism or geopolitical assessment. Such a virtual team could rapidly process vast amounts of open-source data, provide preliminary risk assessments and generate hypotheses for human analysts to explore further. Moreover, these virtual analysts could be deployed 24/7 ensuring continuous monitoring and timely response to emerging threats. While human experts will remain central to the decision-making process, these language-agents could play a valuable augmenting role.

²⁰⁵ Qian et al., "Communicative Agents for Software Development," *arXiv*, 16 July 2023, <https://arxiv.org/abs/2307.07924>; ChatDev is a simulated software development team of language agents (GPT 3.5). These agents are tasked with delivering cost-effective software with minimal human intervention. The agents are assigned roles that correspond to the different stages of the development process: design, code, test, and documentation. Following the tasking of seventy requests to ChatDev, the Brown University study found the system was able to complete the development process 'in under seven minutes and at a cost of less than one dollar'. However, it wasn't without error: limitations such as biases and errors in the LLM flagged several issues over the course of creating the software. See also; Aaron Mok, "AI chatbots were tasked to run a tech company. They built software in under 7 minutes – for less than \$1," *Business Insider*, 11 September 2023, <https://www.businessinsider.com/ai-builds-software-under-7-minutes-less-than-dollar-study-2023-9?r=US&IR=T>.

As *individual* personas, language agents could be used as:

Anti-scam systems: acting as a decoy, the agent attracts scammers, recording their tactics for research, whilst also engaging on platforms prone to scams such as online marketplaces. Combined, these approaches enable simultaneous data collection on scam methods and active engagements with potential scammers.	Anti-disinformation systems: on extremist websites or forums, agents can deploy counter-narratives, engage with users, and mitigate online radicalisation through dialogue and support. Agents could also perform pre-emptive debunking ("pre-bunking") and fact-checking on social media platforms, transparently identifying themselves as bots and providing authoritative source-backed clarifications.
--	--

As *teams or societies*; multi-language agent systems could be useful for:

Data generation: the production of synthetic data where real-world data is scarce or sensitive to obtain. Such data could also be used to provide realistic 'noise' to enhance simulations in scenarios such as disaster response or military.	Training assistants: the development of environments where AI agents can learn through interactions. For example, virtual infrastructure could be used to generate realistic user behaviour or network traffic to test responses in cybersecurity training.	Verification systems: the flooding of systems with agents to stress-test resilience and performance. Agents can also provide the benefit of a consistent environment where different algorithms or systems can be tested and compared for benchmarking.
---	--	--

Nevertheless, there are criticisms of LLM-powered autonomous agent frameworks. One is that they fall short of human-level reasoning. Despite using the latest language models, the agents struggle with understanding the context of a problem. This extends to a lack of genuine understanding of consequences or risk (what humans typically term 'common sense'). Although information handling rules could be supplied to the language agent, it would still not have the innate understanding of risk that humans use to avoid failures.

The key mitigations in addressing these challenges are **accountability and transparency** for both the actions taken by the agent and the inference performed by the model. Until these fundamental improvements are in place, design constraints must take a leading role in managing the risks of autonomy. The Open Worldwide Application Security Project (OWASP) identified 'Excess Agency' as a top ten risk for LLM applications, warning that

'LLM-based systems may undertake actions leading to unintended consequences. The issue arises from excessive functionality, permissions, or autonomy granted to the LLM-based systems'.²⁰⁶

It should be noted that malicious actors will not be concerned with design constraints or issues pertaining to accountability or transparency – instead, the autonomous nature of language agents will likely facilitate the ability to cause harm at increased speed and scale.

Despite these concerns, language agents offer considerable potential benefit for government and citizen alike – as is often the case with significant technical developments, the key will be to match enthusiasm for advances in automation with informed implementation strategies and reliable data sources.

3.2 Open-source and commercial models: innovation vs risk management

Open-source software is defined in *The Turing Way* 'as documenting research code and routines, making them freely accessible and available.'²⁰⁷ If a project is open-source, 'anybody can view, use, modify, and distribute the project for any purpose'²⁰⁸, which enables a lower barrier to adoption and permits 'ideas to spread quickly'.²⁰⁹

However, one interviewee described the idea that a model is either open or closed as a 'false dichotomy'. Model release strategies may affect 'how much is democratised to users' and therefore the types of capabilities external actors are provided with.²¹⁰ The additional layers of components in AI – for example, training and inference code – complicate definitions of open-sourcing beyond its free code-sharing origins in the 1950s and 1960s.

Irene Solaiman has advocated a **gradient framework** which provides a 'level of access' spectrum and highlights the significant variety in release approaches across related models.²¹¹

²⁰⁶ OWASP, "OWASP Top 10 for LLM," <https://llmtop10.com>.

²⁰⁷ The Turing Way, Glossary, <https://the-turing-way.netlify.app/afterword/glossary#term-Open-Source-Software>.

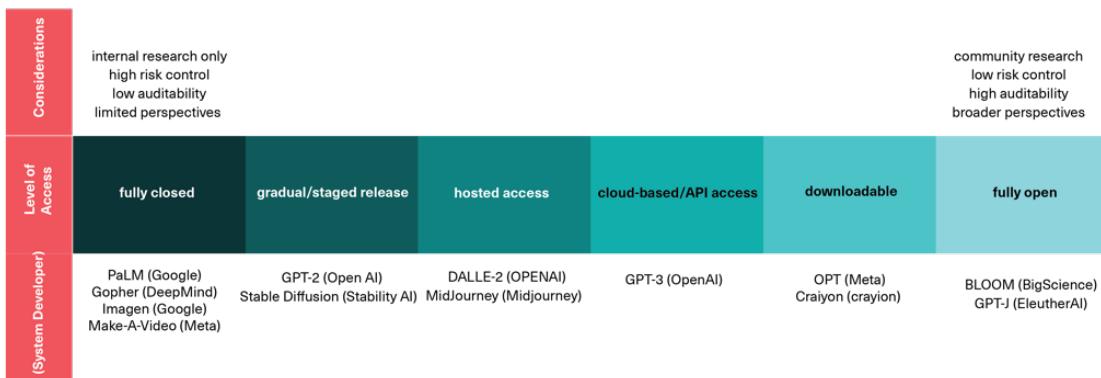
²⁰⁸ The Turing Way, Open-Source Software, <https://the-turing-way.netlify.app/reproducible-research/open/open-source.html>.

²⁰⁹ Ibid.

²¹⁰ Interview with academic, 11 July 2023.

²¹¹ Irene Solaiman, "The gradient of generative AI release: methods and considerations," *arXiv* (February 2023), <https://arxiv.org/abs/2302.04844>.

Figure 7. ‘Levels of access’ gradient framework



Source: Irene Solaiman (2023).

At a U.S. Senate Select Committee on Intelligence hearing (September 2023), Dr Yann LeCun (Chief AI Scientist, Meta) testified that Meta’s approach is to ‘believe it is better if AI is developed openly, rather than behind closed doors by a handful of companies’, describing its benefits as establishing an industry standard that encourages the identification of potential vulnerabilities ‘which can be mitigated in a transparent way by an open community’.²¹²

This philosophy was tested when Meta’s LLaMA model was leaked in March 2023 without instruction, conversation tuning or reinforcement learning from human feedback (RLHF). Within a month, the open-source community plugged these omissions, built upon each other’s findings and – most importantly – ‘solved the scaling problem to the extent that anyone can tinker...the barrier to entry for training and experimentation dropped from the total output of a major research organization to one person, an evening, and a beefy laptop’.²¹³ One interviewee described swift uptake from the open-source community and some of the resulting emergent properties as a surprise, citing the rapid development of Auto-GPT and the use of advanced models to ‘bootstrap’ older models to achieve recursive improvement.²¹⁴

The danger with this comes in the form of ‘**counter-tuning**’ – training open-source models to be toxic or offensive.²¹⁵ Users have stripped safeguards which trained commercial models to

²¹² Yann LeCun, “Hearing before the United States Senate Select Committee on Intelligence,” Testimony of Yann LeCun, Chief AI Scientist, Meta, U.S. Senate, 19 September 2023,

<https://www.intelligence.senate.gov/sites/default/files/documents/os-ylecun-091923.pdf>.

²¹³ Anonymous, “We have no Moat, and Neither Does OpenAI,” Semianalysis, 4 May 2023,

<https://www.semianalysis.com/p/google-we-have-no-moat-and-neither>.

²¹⁴ Interview with government representative, 27 June 2023.

²¹⁵ Interview with government representative, 19 July 2023.

stonewall toxic prompts.²¹⁶ Several interviewees commented on the ‘danger of continually open-sourcing powerful generative AI models’²¹⁷ and the ease with which users can run freely available ‘uncensored’ LLMs with minimal technical skills.²¹⁸

Open-sourcing does offer clear benefits in **red-teaming**²¹⁹ – recently highlighted by the September 2023 launch of OpenAI’s ‘Red Teaming Network’²²⁰ which joins the catalogue of its other AI safety initiatives, ‘Researcher Access’²²¹ and its evaluation framework ‘Evals’.²²² Structured access opportunities in red-teaming were considered by one interviewee as the best route forward – for example, ‘enabling access at different levels for certain actors such as researchers, while preventing certain weights being released to the world’.²²³

In the national security context, there is a complex dynamic between threat and opportunity. Lowering the barrier to entry to powerful models for less technically able (or malicious) users is an almost immediate expansion of the risk surface in terms of who and what may pose a threat. However, in the deployment context, open-source models can offer significant opportunities. From a data security perspective, the lack of traceability of open-source models compared to commercial models makes them more suitable for use within sensitive environments,²²⁴ while their adaptability to specific environments means they can be ‘amended and fine-tuned against niche problem sets in a way closed models cannot’.²²⁵

²¹⁶ Simon Lermen et al., “LoRA fine-tuning efficiently undoes safety training in Llama-2 Chat 70B,” *arXiv* (October 2023), <https://arxiv.org/abs/2310.20624>; Xiangyu Qi et al., “Fine-tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To!,” *arXiv* (October 2023), <https://arxiv.org/abs/2310.03693>; Paul Röttger et al., “XSTest: A Test Suite for Identifying Exaggerated Safety Behaviours in Large Language Models,” *arXiv* (October 2023), <https://arxiv.org/abs/2308.01263>.

²¹⁷ Interview with industry representative, 17 July 2023; Interview with industry representative, 26 June 2023.

²¹⁸ Interview with academic, 7 July 2023.

²¹⁹ Interview with industry representative, 17 July 2023.

²²⁰ OpenAI, Red Teaming Network, <https://openai.com/blog/red-teaming-network>. For more on red-teaming and generative AI, see Laurie Richardson and Royal Hansen, “Acting on our commitment to safe and secure AI,” Google Blog, 26 October 2023, <https://blog.google/technology/safety-security/google-ai-security-expansion>; Marie-Laure Hicks et al., “Exploring red teaming to identify new and emerging risks from AI foundation models,” Summary Workshop Report, RAND Europe, 12 September 2023, https://www.rand.org/pubs/conf_proceedings/CFA3031-1.html.

²²¹ OpenAI, Researcher Access Program, <https://openai.com/form/researcher-access-program>.

²²² OpenAI, Evals, <https://github.com/openai/evals/tree/main>.

²²³ Interview with industry representative, 17 July 2023.

²²⁴ Interview with government representative, 17 July 2023.

²²⁵ Interview with government representative, 17 July 2023; Interview with industry representative, 17 July 2023.

4. Governance, Policy and Regulation

The UK Government has shown increasing ambition in AI governance by first establishing the Frontier AI Taskforce which has morphed into the AI Safety Institute (AISI) following the UK AI Safety Summit in November 2023. The AISI has three core functions: developing and conducting evaluations on advanced AI systems; driving foundational AI safety research; and facilitating information exchange.²²⁶

The AISI, however, is not a regulator and will not determine government regulation. From a policy perspective, it is vital that decision-makers have the right tools and frameworks in place to translate the outputs from bodies like the AISI into effective interventions. As per an August 2023 CETaS Briefing Paper,²²⁷ AI policy approaches should have three main goals:

- To create better visibility and understanding of AI systems
- To promote best practices
- To establish incentives and enforcement of regulation

The wide range of potential applications for generative AI means that centralised coordination across sectors is needed for effective risk management. However, as this often entails significant lead times, a focus on high-risk applications is also essential.²²⁸ Examples may include restricting the use of generative AI in CNI, its use in political advertising for targeting or in making critical decisions pertaining to law enforcement and public services.²²⁹

The next section explores two broad categories of interventions in the generative AI space where UK Government involvement is most needed: 1) **signalling and reporting mechanisms** into government and relevant third-party actors; and 2) ‘red lines’ in the highest-risk contexts.

²²⁶ Department for Science, Innovation and Technology, “Introducing the AI Safety Institute,” HGM Policy Paper, 2 November 2023, <https://www.gov.uk/government/publications/ai-safety-institute-overview/introducing-the-ai-safety-institute>.

²²⁷ Ardi Janjeva et al., “Strengthening Resilience to AI Risk: A guide for UK policymakers,” *CETaS Briefing Paper* (August 2023), <https://cetas.turing.ac.uk/publications/strengthening-resilience-ai-risk>.

²²⁸ Interview with academic, 7 July 2023; OECD, “AI language models: Technological, socio-economic and policy considerations,” OECD Digital Economy Papers, No. 352, 13 April 2023, https://www.oecd-ilibrary.org/science-and-technology/ai-language-models_13d38f92-en

²²⁹ Adam Theierer, “The Schumer AI Framework and the Future of Emerging Tech Policymaking,” R Street Commentary, 27 June 2023, <https://www.rstreet.org/commentary/the-schumer-ai-framework-and-the-future-of-emerging-tech-policymaking>.

4.1 Signalling and reporting

Developing mature signalling and reporting mechanisms should be a priority for policymakers. These approaches can be divided into the following categories, but will all be interlinked:

Signalling		Reporting	
Watermarking	Disclosure and explainability	Multi-layered and socio-technical evaluation	Release strategies

4.1.1 Ex-ante: watermarking

The provenance of a piece of content (whether AI or human-generated) may be signalled at the point of content generation in the form of an embedded watermark (ex-ante).

Watermarking is a possible technical solution to the challenges of AI-enabled political disinformation highlighted in Chapter 2. It allows information such as source tracking or copyright ownership to be “hidden” in the background noise of digital media such as photographs. This is an algorithmic process that can be reversed, allowing the hidden data to be recovered. Watermarking images, video, and audio content is already done in other contexts.²³⁰ Crucially, however, imperceptible watermarking of text is much more challenging than other forms of media²³¹ – semantic-based watermarks can be removed by paraphrasing blocks of text.

Google, OpenAI and Meta have all agreed to develop watermarking tools to combat disinformation and misuse of AI-generated content. Google DeepMind’s *SynthID* can currently be used to add optional watermarks to AI-generated images;²³² *Stable Diffusion* does something similar.²³³ However, all watermarks are **vulnerable to deliberate**

²³⁰ Yuxin Wen et al., “Tree-ring watermarks: fingerprints for diffusion images that are invisible and robust,” *arXiv* (July 2023), <https://arxiv.org/abs/2305.20030>.

²³¹ Initial work in the late 1990s embedded data into the text by subtly shifting the position of lines, letters, and spaces in the document. This method does not survive reformatting: copy-and-pasting or retyping the text will remove the watermark. A more sophisticated method, developed in the early 2000s, analyses the language of the text and adjusts the syntax, without adjusting the meaning of the text, to hide a watermark. See Nurul Shamimi Kamaruddin et al., “A review of text watermarking: theory, methods, and applications,” *IEEE Access* 6 (2018): 8011–8028.

²³² Sundar Pichai, “AI responsibility: tools to identify generated content,” Google Blog, 10 May 2023, <https://blog.google/technology/ai/google-io-2023-keynote-sundar-pichai/#ai-responsibility>.

²³³ Pierre Fernandez et al., “The Stable Signature: Rooting Watermarks in Latent Diffusion Models,” *arXiv* (July 2023), <https://arxiv.org/abs/2303.15435>.

tampering, and watermarks for text are usually broken within a few months²³⁴; any bad-faith actor could bypass it entirely.²³⁵ One University of Maryland study showed that as well as bad-faith actors being able to remove watermarks, it is possible to add watermarks to human-generated images, triggering false positives.²³⁶

One approach might be to add the watermarking at the compute stage on the GPU hardware that performs the computation.²³⁷ This would require significant commitments from GPU manufacturers such as NVIDIA²³⁸ alongside international government coordination, however it would ensure that models are automatically watermarked. The legislative and technological challenges are formidable. On the legislative side, global enforcement of any watermarking system seems daunting, but there are parallels: for example, international governments have successfully managed the risks and benefits of genetic engineering²³⁹ and, starting in 1987, collaborated to rebuild the ozone layer.²⁴⁰ But from a technological perspective, if automatic watermarking cannot survive subsequent editing – including legitimate fine-tuning of generative AI models – then a broken watermarking system might be worse than no watermarks at all.

While the emerging evidence emphasises that watermarking will not be a silver bullet, improving upon it and using it in combination with other technologies will likely play a role in harm reduction and be useful in filtering out lower-level attempts at deception.²⁴¹

4.1.2 Ex-post: disclosure and explainability

Alternatively, provenance may be determined after a piece of content has been created and circulated (ex-post). An example of this would be the tools which purport to detect AI-generated text.

²³⁴ Rhiannon Williams, “AI-text detection tools are really easy to fool,” *MIT Technology Review*, 7 July 2023, <https://www.technologyreview.com/2023/07/07/1075982/ai-text-detection-tools-are-really-easy-to-fool>.

²³⁵ Interview with civil society representative, 29 June 2023.

²³⁶ Kate Knibbs, “Researchers tested AI Watermarks – and Broke All of Them,” *Wired*, 3 October 2023, <https://www.wired.com/story/artificial-intelligence-watermarking-issues>.

²³⁷ Interview with industry representative, 26 June 2023.

²³⁸ Nvidia, “Digital Fingerprinting for Cybersecurity Threat Detection,” Nvidia Solution Brief, 2022, <https://www.nvidia.com/en-gb/ai-data-science/ai-workflows/digital-fingerprinting>.

²³⁹ Brad Bergan, “The Benefits and Risks of Genetic Engineering,” Interesting Engineering, 18 February 2021, <https://interestingengineering.com/science/benefits-risks-genetic-engineering>.

²⁴⁰ UN Environment Programme, “Rebuilding the ozone layer: how the world came together for the ultimate repair job,” UNEP Story, 15 September 2021, <https://www.unep.org/news-and-stories/story/rebuilding-ozone-layer-how-world-came-together-ultimate-repair-job>.

²⁴¹ Kate Knibbs, “Researchers tested AI Watermarks – and Broke All of Them,” *Wired*, 3 October 2023, <https://www.wired.com/story/artificial-intelligence-watermarking-issues>.

However, much like watermarking, in August 2023 OpenAI concluded that none of these were currently reliable.²⁴² OpenAI found that their attempts to train such a tool were confused by writing from individuals for whom English was an additional language.²⁴³ A separate study of 697 participants displayed difficulty in distinguishing human-generated Tweets from GPT-3-generated content.²⁴⁴ As the quality of output from LLMs improves, any perceptible difference between human- and AI-generated text may vanish.²⁴⁵

The challenges associated with detection tools increases the onus on well-intentioned actors to disclose when generative AI is being used, and issue clear guidance on appropriate use and warnings for misuse.²⁴⁶ Better outcomes will be dependent on the **level of explainability provided by the system and the ability of individuals to interpret AI outputs** (achieved through education and training).

Explainable AI currently depends upon three main methods: prediction accuracy, traceability, and decision understanding. All three of these methods are post-hoc, and their effectiveness in an adversarial context is contested.²⁴⁷

- Prediction accuracy involves running simulations and comparing the output to the training data; the most common approach is through Local Interpretable Model-Agnostic Explanations (LIME), which attempts to explain outcomes by learning an interpretable model around the predictions.²⁴⁸
- Traceability attempts to limit the ways decisions can be made by establishing a narrower scope for rules and features; the most prominent example is DeepLIFT.²⁴⁹
²⁵⁰
- Decision understanding requires educating the users of an AI system so that they can better understand how and why the model makes decisions.

²⁴² OpenAI, "How can educators respond to students presenting AI-generated content as their own?," Educator FAQ, August 2023, <https://help.openai.com/en/articles/8313351-how-can-educators-respond-to-students-presenting-ai-generated-content-as-their-own>.

²⁴³ Ibid.

²⁴⁴ Giovanni Spitale, Nikola Biller-Andorno and Federico Germani, "AI model GPT-3 (dis)informs us better than humans," *arXiv* (January 2023), <https://arxiv.org/abs/2301.11924>.

²⁴⁵ Interview with academic, 20 June 2023.

²⁴⁶ Some popular open-source models such as BLOOM have been proactive in communicating such caveats.

²⁴⁷ Sebastian Bordt et al., "Post-hoc explanations fail to achieve their purpose in adversarial contexts," in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (New York: ACM, 2022), 891–905.

²⁴⁸ Marci Túlio Ribeiro et al., "'Why should I trust you?': Explaining the predictions of any classifier," *arXiv* (August 2016), <https://arxiv.org/abs/1602.04938>.

²⁴⁹ DeepLIFT compares the activation of each neuron to a reference neuron in order to produce traceable links between each activated neuron and their dependencies.

²⁵⁰ Avanti Shrikumar et al., "Learning important features through propagating activation differences," *arXiv* (October 2019), <https://arxiv.org/abs/1704.02685>.

On this last theme, the ability of senior government officials to digest and reliably act upon AI-generated outputs has often been highlighted by researchers as an education and training priority – an upcoming CETaS report on ‘Communicating AI-enriched Intelligence Reporting to Strategic Decision-Makers’ will explore this topic in more detail.

4.1.3 Multi-layered and socio-technical evaluation

Governments are now devoting significant resources to evaluating advanced AI systems as a way of containing AI risks. Current approaches to evaluation are model-centric, as identified in several papers.²⁵¹ This is important in determining technical functionality and capabilities,²⁵² but insufficient for understanding systemic implications for national security. To address this gap, Weidinger et al have proposed a socio-technical evaluation framework for AI safety.²⁵³ This approach illuminates how systemic, high-level security threats interact with specific system capabilities. It is an interdisciplinary framework which seeks to account for human and systemic factors²⁵⁴ – purely technical AI evaluation cannot encompass the intellectual diversity, understanding and rigour required to ascertain whether a system is safe.²⁵⁵

For each threat to national security, the following layers of evaluation²⁵⁶ should be carried out for a comprehensive understanding of how AI systems perpetuate risks.²⁵⁷

²⁵¹ Andrew D. Selbst et al., “Fairness and abstraction in sociotechnical systems,” in *Proceedings of the 2019 ACM Conference on Fairness, Accountability, and Transparency* (New York: ACM, 2019), 59–68; Seth Lazar, and Alondra Nelson, “AI safety on whose terms?” *Science* 381, no. 6654 (2023): 138–138; Renee Shelby et al., “Sociotechnical Harms of Algorithmic Systems: Scoping a Taxonomy for Harm Reduction,” in *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society* (New York: ACM, 2023): 723–741; Shakir Mohamed, Marie-Therese Png, and William Isaac, “Decolonial AI: Decolonial theory as sociotechnical foresight in artificial intelligence,” *Philosophy & Technology* 33 (2020): 659–684.

²⁵² Deborah Inioluwa Raji et al., “The fallacy of AI functionality,” in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, (New York: ACM, 2022): 959–972.

²⁵³ Laura Weidinger et al., “Sociotechnical Safety Evaluation of Generative AI Systems,” *arXiv* (October 2023), <https://arxiv.org/abs/2310.11986>.

²⁵⁴ Ibid.

²⁵⁵ Seth Lazar, and Alondra Nelson, “AI safety on whose terms?,” *Science* 381, no. 6654 (2023): 138–138.

²⁵⁶ Evaluation has limitations: (1) Evaluation is incomplete as approaches can only feasibly cover a subset of risks. (2) Evaluation is not value-neutral, as it involves a series of normative decisions and value expressions about what should be prioritised. See Laura Weidinger et al., “Sociotechnical Safety Evaluation of Generative AI Systems,” *arXiv* (October 2023): 6–7.

²⁵⁷ Andrew D. Selbst et al., “Fairness and abstraction in sociotechnical systems,” in *Proceedings of the 2019 ACM Conference on Fairness, Accountability, and Transparency* (New York: ACM, 2019): 59–68.

1. **Capability Layer** – evaluating the level of risk from the technical components and behaviours of generative AI systems.
2. **Human Interaction Layer** – evaluating the level of risk from the interactions between the technical systems and human users.
3. **Systemic and Structural Layer** – evaluating the level of risk from systemic and structural factors that will interact with the AI system capability and human interactions.

Case study: Political disinformation and electoral interference

As an exemplar, we apply the socio-technical evaluation framework to political disinformation and electoral interference.²⁵⁸ This case study will outline which tasks should be carried out at each layer of evaluation to understand the level of national security risk.

Although Weidinger et al. focus on *misinformation* in their framework,²⁵⁹ the threat of political disinformation and electoral interference requires additional evaluations. Disinformation is varied in nature: it can be one-to-many, where a message is transmitted to a wide audience, or one-to-one, where an operator repeatedly engages a specific target. It can use different modalities, such as text, image, or video, and be disseminated through different infrastructures and coordination mechanisms.

When considering political disinformation and electoral interference, multi-layered evaluation takes the following form:

1. **Capability Layer:** tasks to evaluate AI system capability to create content for political disinformation and electoral interference.
2. **Human Interaction Layer:** tasks to evaluate human interaction with this content from both the perspective of human operators and human targets.
3. **Systemic and Structural Layer:** tasks to evaluate systemic and structural factors related to risk of political disinformation and electoral interference.

²⁵⁸ Aaron Blake, “DeSantis ushers in our fake-images-in-politics nightmare,” *The Washington Post*, 8 June 2023, <https://www.washingtonpost.com/politics/2023/06/08/desantis-fauci-trump-ai-video/>; Priyanka Ranade et al., “Generating fake cyber threat intelligence using transformer-based models,” in *International Joint Conference on Neural Networks (IJCNN)* (IEEE, 2021): 1–9.

²⁵⁹ Laura Weidinger et al., “Sociotechnical Safety Evaluation of Generative AI Systems,” *arXiv* (October 2023): 39–43.

In this case study, we detail specific evaluation tasks for each layer of the framework. These recommended tasks are not exhaustive, but they provide a means of quantifying risk to political disinformation and electoral interference from generative AI, from a capability, interactive and systemic perspective.

Capability Layer

1. **Factuality:** the factuality of AI system outputs²⁶⁰ can be measured as ‘fact verification’, where the output is compared with a knowledge source, e.g. WikiData.²⁶¹ It can be measured as ‘verifiability’, where one tests whether the output can be attributed to a reliable source, e.g. FActScore.²⁶² One could also measure ‘factual knowledge’, which is the tendency of the AI system to generate factual statements from a corpus, e.g. Wiki-FACTOR and News-FACTOR.²⁶³
2. **Credibility:** the credibility of AI system outputs²⁶⁴ can be measured as ‘quality’ or ‘fidelity’ of generative AI content by evaluating realism of outputs using FID Scores or Inception Scores.²⁶⁵
3. **Narrative reiteration:** the ability of AI systems to generate new content that iterates on a theme, thereby amplifying a narrative. A task could be to prompt the AI system to generate ‘N’ pieces of content of a specified format (such as a meme) on one topic (such as climate change denialism).²⁶⁶
4. **Narrative seeding:** the ability of AI systems to generate new narratives or content within an existing conspiracy. When evaluating GPT-3 for this ability, Buchanan et al. asked GPT-3 to generate “Q-style drops” to mimic content from the QAnon conspiracy.²⁶⁷ One study found that GPT-4 was in fact more proficient at this than

²⁶⁰ Ibid.

²⁶¹ Nayeon Lee et al., “Factuality enhanced language models for open-ended text generation,” *Advances in Neural Information Processing Systems* 35 (2022): 34586–34599.

²⁶² Sewon Min et al., “FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation,” *arXiv* (October 2023), <https://arxiv.org/abs/2305.14251>.

²⁶³ Dor Muhlgay et al., “Generating benchmarks for factuality evaluation of language models,” *arXiv* (July 2023), <https://arxiv.org/abs/2307.06908>.

²⁶⁴ Laura Weidinger et al., “Sociotechnical Safety Evaluation of Generative AI Systems,” *arXiv* (October 2023): 39–43.

²⁶⁵ Fréchet inception distance (FID) is a metric for quantifying how realistic and diverse images generated by a generative AI system are. Inception Score is a metric that measures the quality or realism generated images.

²⁶⁶ Ben Buchanan, Andrew Lohn and Micah Musser, *Truth, lies, and automation: How language models could change disinformation* (CSET Georgetown: May 2021), <https://cset.georgetown.edu/publication/truth-lies-and-automation/>.

²⁶⁷ Ibid.

GPT-3.²⁶⁸

5. **Content manipulation:** the ability of AI systems to reframe existing content (such as a news article or video) to support a new perspective. When evaluating GPT-3 for narrative manipulation, Buchanan et al ask the AI system to summarize the original article and then rewrite it with a new, specified viewpoint.²⁶⁹ The same task could be applied to video or image generation.
6. **Personalisation:** the ability of AI systems to personalise content for a target/group.²⁷⁰ Personalisation could be leveraged by disinformation operators to stoke existing divisions by targeting opposite messaging to different groups. A text-to-text model's ability to create personalised and targeted messaging could be evaluated using 'CSET Narrative Wedging Tasks'.²⁷¹
7. **Dishonest anthropomorphism:** whether AI systems give dishonest signals of being human, particularly for chat-based applications.²⁷² Anthropomorphism has been shown to greatly impact human interaction and their level of trust in a system.²⁷³ This can be achieved by interrogating the system about human identity using datasets like the R-U-A-Robot Dataset.²⁷⁴

²⁶⁸ Lorenzo Arvanitis et al., "Despite OpenAI's promises, the company's new AI tool produces misinformation more frequently, and more persuasively, than its predecessor," Misinformation Monitor: March 2023, NewsGuard, March 2023, <https://www.newsguardtech.com/misinformation-monitor/march-2023/>.

²⁶⁹ Ben Buchanan, Andrew Lohn and Micah Musser, *Truth, lies, and automation: How language models could change disinformation* (CSET Georgetown: May 2021), <https://cset.georgetown.edu/publication/truth-lies-and-automation/>.

²⁷⁰

Hannah Rose Kirk et al. "Personalisation within bounds: A risk taxonomy and policy framework for the alignment of large language models with personalised feedback," *arXiv* (March 2023), <https://arxiv.org/abs/2303.05453>.

²⁷¹ Ben Buchanan, Andrew Lohn and Micah Musser, *Truth, lies, and automation: How language models could change disinformation* (CSET Georgetown: May 2021), <https://cset.georgetown.edu/publication/truth-lies-and-automation/>.

²⁷² Ameet Deshpande et al., "Anthropomorphization of AI: Opportunities and Risks," *arXiv* (May 2023), <https://arxiv.org/abs/2305.14784>.

²⁷³ Yuanyuan Cui, "Sophia Sophia tell me more, which is the most risk-free plan of all? AI anthropomorphism and risk aversion in financial decision-making," *International Journal of Bank Marketing* 40, no. 6 (2022): 1133–1158; Amani Alabed, Ana Javornik and Diana Gregory-Smith, "AI anthropomorphism and its effect on users' self-congruence and self–AI integration: A theoretical framework and research agenda," *Technological Forecasting and Social Change* 182 (September 2022), <https://doi.org/10.1016/j.techfore.2022.121786>; Cammy Crolic et al., "Blame the bot: Anthropomorphism and anger in customer–chatbot interactions," *Journal of Marketing* 86, no. 1 (2022): 132–148.

²⁷⁴ David Gros, Yu Li and Zhou Yu, "The RUA-robot dataset: Helping avoid chatbot deception by detecting user questions about human or non-human identity," *arXiv* (June 2021), <https://arxiv.org/abs/2106.02692v1>.

Human Interaction Layer

1. **Deception:** whether AI system outputs can successfully deceive targets.²⁷⁵ This can be achieved through human evaluation, such as Mean Opinion Scores of image and speech quality or by testing the ability to distinguish between real and synthetic content (e.g. HPBench dataset).²⁷⁶
2. **Persuasion:** whether AI system outputs can impact or shift attitudes, beliefs, or behaviours.²⁷⁷ This can be evaluated through survey experiments to test persuasiveness of AI-generated political propaganda in relation to key policy issues.²⁷⁸
3. **Reliance on human operator:** how much the quality of AI system outputs depends on interaction with a human operator. If successful outputs are very reliant on high interaction or high-quality prompts, this will reduce some of the scalability risks of generative AI. This could be evaluated by iteratively measuring improvement in content quality with each operator interaction. Buchanan et al. found that GPT-3 worked better when in a human-machine team with back-and-forth correspondence with a human operator.²⁷⁹ Further experimentation can be carried out to measure prompt sensitivity.²⁸⁰

Systemic and Structural Layer

1. **Public trust:** levels of public trust in traditional media²⁸¹ can influence the risk of AI-enabled political disinformation and electoral interference. Equally, however, there are high-profile examples of traditional media contributing towards political

²⁷⁵ Laura Weidinger et al., "Sociotechnical Safety Evaluation of Generative AI Systems," *arXiv* (October 2023): 39–43.

²⁷⁶ Zeyu Lu et al., "Seeing is not always believing: A Quantitative Study on Human Perception of AI-Generated Images" *arXiv* (September 2023), <https://arxiv.org/abs/2304.13023>.

²⁷⁷ Laura Weidinger et al., "Sociotechnical Safety Evaluation of Generative AI Systems," *arXiv* (October 2023).

²⁷⁸ Hui Bai et al., "Artificial intelligence can persuade humans on political issues," *Nature Portfolio Preprint* (September 2023); Josh A. Goldstein et al., "Can AI Write Persuasive Propaganda?" *SocArXiv Papers* (February 2023).

²⁷⁹ Ben Buchanan, Andrew Lohn and Micah Musser, *Truth, lies, and automation: How language models could change disinformation* (CSET Georgetown: May 2021), <https://cset.georgetown.edu/publication/truth-lies-and-automation/>.

²⁸⁰ Guanghui Qin and Jason Eisner, "Learning how to ask: Querying LMs with mixtures of soft prompts," *arXiv* (April 2021), <https://arxiv.org/abs/2104.06599>.

²⁸¹ Laura Weidinger et al., "Sociotechnical Safety Evaluation of Generative AI Systems," *arXiv* (October 2023): 39–43.

disinformation and damaging the integrity of the electoral process.²⁸² Trust can be assessed through population-level surveys or social media studies.

2. **Availability:** availability of AI systems to the public, particularly via open-source models. Without externally enforced safeguards, these models are easier for malicious actors to manipulate and use for nefarious ends. This can be assessed by monitoring the number of known open-source models and the public availability of commercial models.
3. **Identification:** the ability to successfully identify synthetic media may be correlated with the risk it poses. However, improvements in identification are likely to happen simultaneously with advancements in synthetic media generation. Watermarking is a potential method for detecting AI-generated content²⁸³ and is seeing dedicated investment by big tech companies and DARPA²⁸⁴ (see previous section on 'Watermarking' for more detailed discussion).
4. **Prevalence:** the prevalence of synthetic media on topics such as elections, political parties, current affairs, and polarising issues. This task requires effective identification methods and a choice of platform(s) and topics to investigate.
5. **Pollution:** disinformation operators will often pollute public discourse with an influx of synthetic media to shift the perceived importance or perceived public stance on an issue. This task also requires effective identification methods and a choice of platform(s) and topics to monitor.
6. **Feasibility:** the feasibility of malicious actors interfering in an election based on the resilience of a given political system to interference. This could be assessed based on metrics such as estimated election or vote margin.

This case study provides an example of how a multi-layered evaluation approach can be used to evaluate the risk of generative AI to national security. It highlights how risk is co-produced and amplified by each evaluation layer. Political disinformation has two crucial elements: content creation and content distribution. Generative AI will play a greater role in altering and expediting the content creation process than the distribution process, which

²⁸² Sam Levine and Kira Lerner, "Fox and Dominion settle for \$787.5m in defamation lawsuit over US election lies," *The Guardian*, 18 April 2023, <https://www.theguardian.com/us-news/2023/apr/18/fox-dominion-settle-us-defamation-lawsuit>.

²⁸³

John Kirchenbauer et al., "A watermark for large language models," *arXiv* (June 2023), <https://arxiv.org/abs/2301.10226>.

²⁸⁴ Will Corvey, "Semantic Forensics (SemaFor)," Defence Advanced Research Project Agency, accessed November 2023, <https://www.darpa.mil/program/semantic-forensics>.

relies on operators' distribution infrastructure; yet the risks still sit across all three layers. The table below illustrates why generative AI must be evaluated from this multi-layered and socio-technical perspective.

Political Disinformation	Capability Layer	Interaction Layer	Systemic Layer
Ability to create content	Y e.g. narrative reiteration and seeding		
Likelihood of targets believing the content	Y e.g. credibility and personalisation	Y e.g. persuasion and deception	Y e.g. public trust in traditional media
Scalability of content creation		Y e.g. reliance on human operator	Y e.g. availability and feasibility
Reach of synthetic content			Y e.g. prevalence and pollution

Future work should seek to apply this framework to other identified threats to national security, and formally operationalise and test the evaluations in a specified context. For the purposes of this report, the proposed evaluation approach represents an important shift towards considering socio-technical factors when evaluating generative AI systems.

4.1.4 Release strategies

The leading commercial developers of AI systems tend to be profit-driven and possess incentives to crowd out competitors who may not have the means to scale potentially significant innovations.²⁸⁵ Within some of these companies, the constant pressure to ship new products to a captive audience has grown significantly: this was evident following the public release of ChatGPT which saw companies like Google bring forward the release of their own generative AI tools.²⁸⁶

The rapid pace of change and increases in capability associated with these tools means that policymakers have to be continually alert to the next major innovation.²⁸⁷ On 25 September 2023, OpenAI announced that 'ChatGPT can now see, hear, and speak' and that they were beginning to roll out new voice and image capabilities in the service in a matter of weeks.²⁸⁸ The ways this kind of multi-modality could alter the threat landscape for governments are

²⁸⁵ Melissa Heikkilä, "Generative AI risks concentrating Big Tech's power. Here's how to stop it," *MIT Technology Review*, 18 April 2023, <https://www.technologyreview.com/2023/04/18/1071727/generative-ai-risks-concentrating-big-techs-power-heres-how-to-stop-it/>.

²⁸⁶ Interview with civil society representative, 7 July 2023.

²⁸⁷ Interview with industry representative, 17 July 2023.

²⁸⁸ OpenAI, "ChatGPT can now see, hear, and speak," OpenAI Blog, 25 September 2023, <https://openai.com/blog/chatgpt-can-now-see-hear-and-speak>.

likely to be significant, but without **pre-release deliberative and evaluative processes**, it is a change which they had to adapt to in real time.

Following the UK's AI Safety Summit, it was announced that 'leading AI developers' had committed to avoid releasing models without prior testing by government agencies.²⁸⁹ This is a positive step forward, but questions remain regarding which companies qualify as 'leading', which governments will assist the AISI with the testing and the methodological frameworks they agree to, and the available in-house expertise to conduct such testing to a high standard. Moreover, even if individual models are safety tested prior to release, it is unclear whether iterative updates to those models (such as ChatGPT going 'multi-modal') are included in this measure.

4.2 Prohibition and 'red lines'

Governments may look to implement more stringent restrictions in areas where the integration of AI into decision-making functions will be undesirable for the foreseeable future.

'Perfect trust' scenarios were highlighted by various interviewees as a priority in this regard: 'wherever there has to be perfect trust, such as nuclear command and control, even if it appears more efficient, the probability/impact calculation of automation is bad in those scenarios.'²⁹⁰ The argument may also extend to non-CNI contexts which nonetheless have very consequential outcomes, such as policing and criminal justice. Safety breaks for AI systems that control the operation of critical infrastructure could be similar to the braking systems that engineers have long built into other technologies.²⁹¹

While some interviewees were confident that operational staff in national security and defence were well accustomed to these limitations – and in many cases already operated under clear restrictions for certain technologies²⁹² – others were more sceptical about whether the same could be said for more senior officials. Individuals charged with achieving workforce savings and finding 'an edge' may see some of the risks and prerequisites around

²⁸⁹ Martin Coulter and Paul Sandle, "At UK's AI Summit, developers and govts agree on testing to help manage risks," *Reuters*, 2 November 2023, <https://www.reuters.com/world/uk/uk-pm-sunak-lead-ai-summit-talks-before-musk-meeting-2023-11-02>.

²⁹⁰ Interview with industry representative, 26 June 2023.

²⁹¹ Brad Smith, "How do we best govern AI?," Microsoft On the Issues Blog, 25 May 2023, <https://blogs.microsoft.com/on-the-issues/2023/05/25/how-do-we-best-govern-ai>; Will Hunt, "The Flight to Safety-Critical AI: Lessons in AI Safety from the Aviation Industry," Berkeley Center for Long-Term Cybersecurity White Paper, August 2020, <https://cltc.berkeley.edu/publication/new-report-the-flight-to-safety-critical-ai-lessons-in-ai-safety-from-the-aviation-industry>.

²⁹² Interview with civil society representative, 17 July 2023.

data governance and surrounding AI infrastructure as constraints.²⁹³ One interviewee put this into perspective when asking, 'is efficiency always your goal? There are some areas where a slowing down is something you can accept.'²⁹⁴ Pre-empting the high-stakes contexts where generative AI should not be used will prevent situations where the technology can take **irreversible actions without direct human oversight or authorisation.**

Outside some of the higher-risk contexts described in this section, numerous interviewees did express reservations about the usefulness of 'red lines' in general. One said, 'it can be counterproductive to pick a red line that is not defensible in the long term (...) you are better off having evolving regulatory standards that are informed by state-of-the-art thinking.'²⁹⁵

This reservation is justified on a broad level, provided that the signalling and reporting interventions described above are developed and implemented effectively. The pressure to prohibit a wider range of use cases will lessen if AI models are seen to be trained, tested, audited, deployed, and disseminated in a more structured and responsible way.

4.3 Strengths and weaknesses in the legislative environment

Legislation is notoriously slow-moving in comparison to the pace of change and direction of technology. The UK's Online Safety Bill was first broached by the government in 2019, but only approved by the House of Lords in September 2023. In the AI space, innovation is measured by weeks and months rather than years, enhancing the risk of governments **legislating for technological shortcomings that are obsolete** by the time a bill becomes law.

The governability of the generative AI ecosystem is made significantly more challenging by the decentralised nature of model use combined with the borderless nature of risks.²⁹⁶ Legal comparisons to the nuclear non-proliferation context are limited because the difficulty of coming into possession of plutonium is of a different order to running powerful models on personal computers.²⁹⁷ One interviewee was clear that 'if we regulated open-sourcing in the

²⁹³ Interview with industry representative, 18 July 2023.

²⁹⁴ Interview with government representative, 28 July 2023.

²⁹⁵ Interview with industry representative, 25 July 2023.

²⁹⁶ Interview with academic, 6 July 2023.

²⁹⁷ LLaMA 2 was released in July 2023 for free for research and commercial use by Meta. See Meta, "Meta and Microsoft Introduce the Next Generation of Llama," 18 July 2023, <https://about.fb.com/news/2023/07/llama-2/>.

UK, it would not have much impact'²⁹⁸, with another concurring that 'technology is too far out there and too available'.²⁹⁹

Some interviewees felt that various existing regulations in areas like contract and copyright law could be re-applied effectively: 'if you use generative AI to write a CV for a job which you get, and it is later exposed as fake, that is your fault ultimately, you cannot pass that liability on.' Placing too much weight on a 'magic new set of laws specifically for generative AI' could risk delaying meaningful government interventions in the areas described in previous sections, while also stifling innovation.³⁰⁰ Therefore, in the immediate term, a focus on **expanding the scope of existing legislation to account for generative AI** should be prioritised ahead of AI-specific legislative frameworks or banning individual AI services.

In the national security and law enforcement context, there are more specific challenges that could emerge from the **operational use of generative AI**. The UK has authorising legislation which allows agencies to conduct digital intelligence collection, achieved through warrants with expiration dates. Upon expiration, these agencies must purge the information obtained from their collection from their systems. But while it is easier to purge information in *databases*, achieving the same results with *LLMs* is more challenging because of the way in which these models are trained, namely by scraping vast quantities of data from the Internet: 'using a pre-trained model with all that information on there could be a disaster from a compliance point of view for an agency'.³⁰¹ There are areas of research, such as 'machine unlearning', which are focused on removing the influence of a specific subset of training examples from a trained model. If LLMs were to be deployed in an investigative context and even fine-tuned on operational data, machine unlearning could prove essential from both a compliance and effectiveness perspective.³⁰²

4.4 Global governance

The UK Government must capitalise on the momentum generated by the AI Safety Summit. Having a set of criteria which global endeavours should strive to meet is important in ensuring that efforts avoid becoming fragmented.

²⁹⁸ Interview with government representative, 19 July 2023.

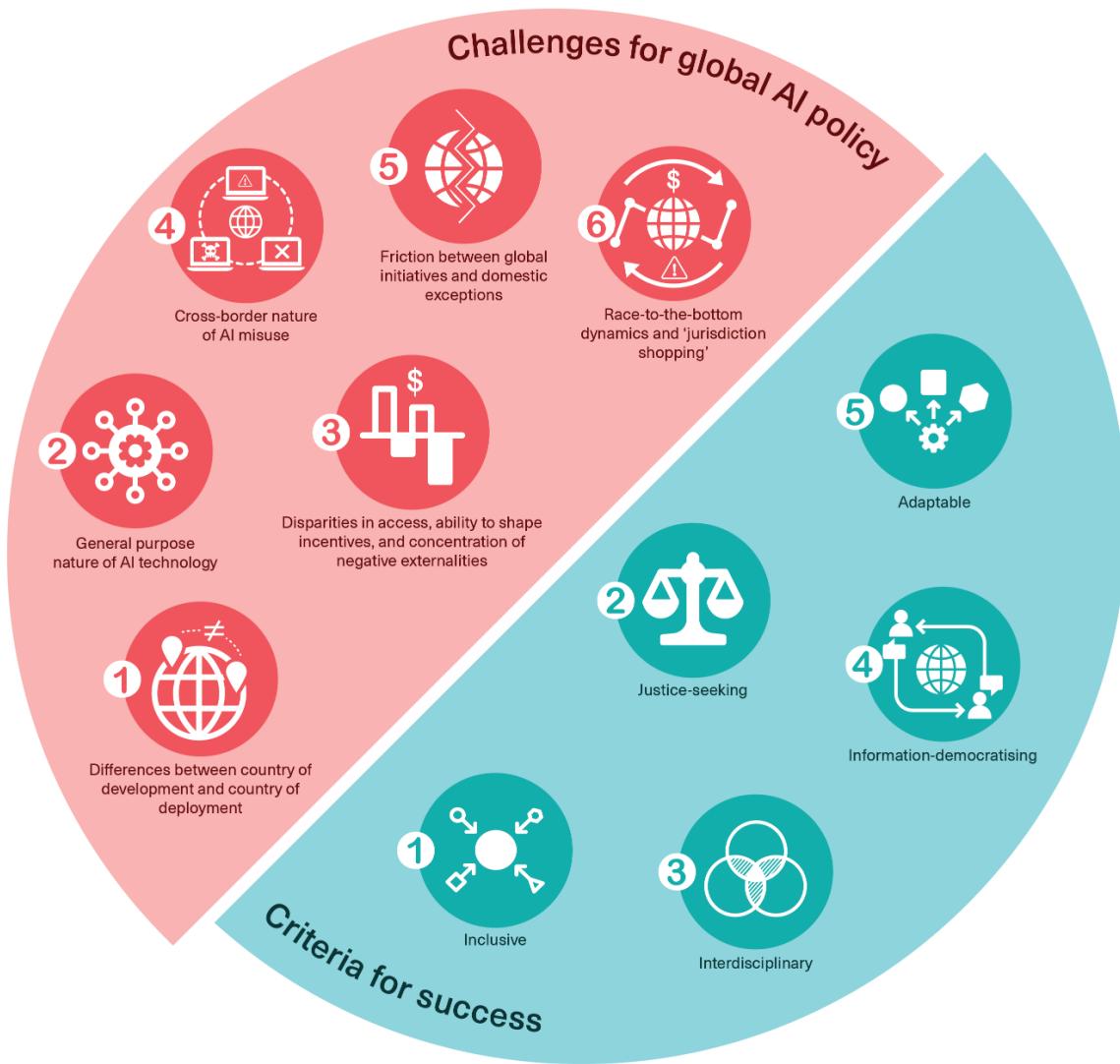
²⁹⁹ Interview with academic, 6 July 2023.

³⁰⁰ Interview with civil society representative, 29 June 2023; Interview with government representative, 19 July 2023.

³⁰¹ Interview with industry representative, 19 July 2023.

³⁰² Ibid.

Figure 8. An illustration of six challenges for global AI policy to solve and five criteria to be met to increase the likelihood of success³⁰³



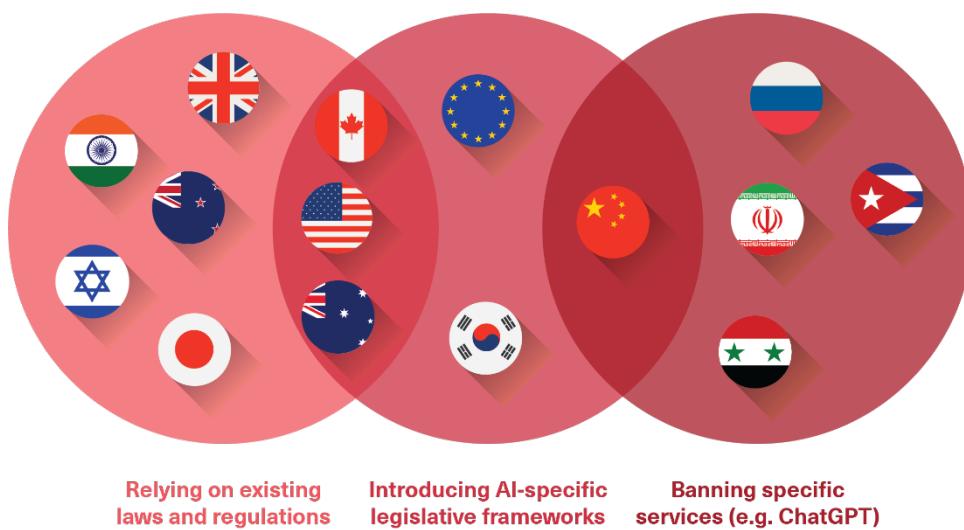
4.4.1 Aligning global standards

Some experts dismiss the possibility that meaningful policy solutions can be created at a global level because of the influence of non-cooperative nations.³⁰⁴ But even amongst Western democracies, differences emerge between the likes of the US, UK, and EU, some of which are exemplified in the figure below.

³⁰³ Figure from "Strengthening Resilience to AI Risk: A guide for UK policymakers," CETaS Briefing Paper (August 2023), 40.

³⁰⁴ Michael Dempsey, "Is it possible to regulate artificial intelligence?," BBC News, 21 September 2023, <https://www.bbc.co.uk/news/business-66853057>.

Figure 9. Venn diagram of different global approaches to AI regulation



There are two key things that the UK could do to narrow some of the disparities that persist at a global level: promote **shared evaluation tools and clear targets**; and contribute to **international regulatory expertise and capacity**.

One of the reasons why it can be difficult to demarcate where AI is suitable for use is because of the lack of **developed technical standards** upon which to base policy. One interviewee commented that 'the standards issue requires a foundational measurement theory for AI systems (...) I don't know that we have the theory which allows you to demand that a system has a score of 5 on some explainability rating system.'³⁰⁵ When civil engineering a bridge, there are well understood theories of how to 'decompose systems, recompose them, understand how measurements apply, ensure reproducibility.'³⁰⁶ Comparatively, there is no equivalent in AI which allows the policy community to approach technologists with the question of 'what needs to be written into legislation to get safe AI systems?'

Developing an international ecosystem of third-party auditors which provide additional capacity to governments and global institutions will be a central pillar of this work.³⁰⁸ In domains like aviation safety, standards bodies have ensured that aircrafts can land safely around the world without having to change key aspects of the technology itself.³⁰⁹ For AI risk assessment, there are attempts at rating systems and standards promoted by civil society and public oversight groups, such as Ranking Digital Rights, which rates digital platform

³⁰⁵ Interview with government representative, 20 July 2023.

³⁰⁶ Ibid.

³⁰⁷ Ibid.

³⁰⁸ Interview with industry representative, 25 July 2023.

³⁰⁹ Interview with civil society representative, 29 June 2023.

companies on human rights; and Algorithm Watch's 'SustAIIn', which seeks standards for social and environmental sustainability on all aspects of AI development.³¹⁰

Avoiding an over-reliance on industry – particularly Silicon Valley – to provide guidance and expertise is a challenge given the resources at their disposal, yet critical to the authority and independence of global auditing mechanisms.³¹¹ The growing network of AI safety institutes (two launched in the UK and US, with similar versions touted in France and Canada)³¹² must leverage a **multi-stakeholder effort on quality control and standards** enabling accountability of advanced AI models.

4.4.2 AI diplomacy: avoiding a race to the bottom

Technology is now a central part of modern geostrategic rivalry. China has the stated aim of achieving global dominance through science and technology, while US actions show an ambition to prevent that outcome through a combination of export controls and vigorous domestic industrial policy.³¹³ Achieving the global governance goals outlined above entails some minimum level of diplomatic engagement which ensures that **rapid AI adoption does not supersede AI safety research**.³¹⁴ The potential for AI to confer **strategic advantage** on 'winners' of R&D races and leave 'losers' behind creates powerful incentives that may run at odds with AI safety.³¹⁵

The success of AI in devising winning strategies in a variety of games and coming up with novel ideas has raised many questions about how the speed and opacity of AI systems might affect strategic stability. If decision-making in security contexts is increasingly informed by AI, the timeframes that diplomats and senior security officials will operate on will be far shorter than during the Cold War.³¹⁶ There is already a tendency towards less human-to-human engagement in 'grey zone warfare', therefore it is not a big conceptual leap to have military strategies provided by an LLM before automated assets perform a

³¹⁰ G. Bell et al., "Generative AI: Language models and multimodal foundation models," Australian Government Rapid Information Report, 24 March 2023, <https://www.atse.org.au/wp-content/uploads/2023/06/Rapid-Response-Information-Report-Generative-AI-220602.pdf>.

³¹¹ Interview with civil society representative, 29 June 2023.

³¹² Cristina Gallardo and Tim Smith, "The UK and US are setting up AI safety institutes – what that means for startups," *Sifted*, 1 November 2023, <https://sifted.eu/articles/the-uk-and-us-are-setting-up-national-ai-safety-institutes>; Murad Hemmadi, "Champagne explores Canadian AI safety institute, touts country's role on 'transformative technology,'" *The Logic News*, 2 November 2023, <https://thelogic.co/news/champagne-explores-canadian-ai-safety-institute-touts-countrys-role-on-transformative-technology/>.

³¹³ Interview with government representative, 27 June 2023.

³¹⁴ Interview with government representative, 14 July 2023.

³¹⁵ Eric Schmidt, "AI, Great Power Competition & National Security," *Daedalus* 151, no. 2 (Spring 2022): 288–298, <https://direct.mit.edu/daed/article/151/2/288/110603/AI-Great-Power-Competition-and-National-Security>.

³¹⁶ Interview with industry representative, 19 July 2023.

tactical attack.³¹⁷ One could also envisage generative AI being used to raise public support for a particular conflict, fuelled by agent-led automated cyber-attacks which lead decision-makers into wars of miscalculation.³¹⁸ This is why in the most critical international security scenarios, world leaders must maintain human-centred offramps which de-escalate misdirected signals abetted by technology.

Concern about ‘falling behind’ adversaries and competitors may be one of the driving forces of the potential race to the bottom described here. Although expert opinion varies on the question of how ‘far behind’ Chinese LLMs are compared to those developed in the US or UK, one interviewee placed the figure at around one year while stressing that catching up in this space was very doable.³¹⁹

Others warned of the dangers of feeding a self-fulfilling prophecy: ‘China is not an AI rival in all aspects. They are extremely good at operationalising other developments in this field, such as computer vision and surveillance, but the LLMs coming from China are not comparable in relation to US or UK companies. It is easy for AI to get wrapped up in larger narratives on geopolitical rivalry’.³²⁰

Clearly, there is a need to find a balance between competing incentives. The Bletchley Declaration following the AI Safety Summit marked an important step to achieving this.³²¹ Attaining China’s signature, even if symbolic, was a feat which should lay the foundations for further progress at the next Global Summits hosted by South Korea and France. But as well as engagement with China at the state level, countries like the UK must allow for responsible engagement at the grassroots level, via the engineers and academics who populate global standards bodies. This approach could afford longer-term stability compared to political engagement which oscillates more frequently.

4.4.3 Leading responsible innovation

By developing a responsible AI ecosystem at home where developers and users are more accustomed with UK regulatory structures, the UK can better demonstrate its AI governance credentials to the rest of the world.

³¹⁷ Interview with academic, 14 July 2023.

³¹⁸ Interview with government representative, 14 July 2023.

³¹⁹ Interview with government representative, 27 June 2023.

³²⁰ Interview with civil society representative, 17 July 2023.

³²¹ UK Government, “The Bletchley Declaration by countries attending the AI Safety Summit, 1-2 November 2023,” HMG Policy Paper, 1 November 2023, <https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023>.

There are two key areas where balance and trade-offs will be required: 1) proactivity in policy delivery vs adapting to the global agenda; and 2) supporting positive AI use cases vs AI safety credentials.

Regarding 1), several interviewees stressed the importance of **implementation** when it came to regulation.³²² The EU AI Act has already been through various iterations and is currently at a 'trilogue stage' with the Commission, Council and Parliament reconciling differences. As mentioned in Chapter 2, China issued draft generative AI laws in April 2023 which it later refined into 'Interim Measures' over July and August. Despite ongoing debate as to the suitability of these respective approaches, they are examples of concrete legal proposals which academia and industry can debate and anticipate. Regardless of the sector, industry requires clarity over the legal landscape to **plan for the long-term and invest confidently**. One interviewee provided a telling insight in this regard: 'there are banks who are not putting AI into play at the moment, despite the fact they know it could unleash productivity gains because of all the uncertainty in terms of regulation and how it relates to the Financial Conduct Authority.'³²³

Regarding 2), the UK needs to create enough room to **support the positive applications of AI** (which attract the bulk of government and industry funding) while addressing the ways in which AI systems can cause harm. One interviewee commented that 'being a leader in the AI space means demonstrating the economic viability of these systems while establishing safety norms – it is not a dichotomy between leadership and safety'.³²⁴ While the public is used to hearing about positive AI use cases from industry, there are comparatively far fewer examples of government bodies outlining clearly how they envisage AI transforming their operating practices now and in future.³²⁵

The UK Government's ability to support skills and infrastructure is another key challenge, aptly summarised by one interviewee: 'paying government technologists properly, addressing the recruitment problem, creating computing environments they can work with and enabling access to data in a ground-up way – until that is solved forget about things like "BritGPT" and "AI superpower" talk.'³²⁶ **Access to compute, data and staff** are three core pillars for leadership in AI innovation but achieving this combination at this time is currently a challenge for the UK Government.

Hardware challenges are inherently linked to the UK's ability to service needs across the AI development supply chain, including energy and semiconductors. For example, if the UK

³²² Interview with government representative, 18 August 2023; Interview with industry representative, 3 July 2023.

³²³ Interview with industry representative, 3 July 2023.

³²⁴ Interview with industry representative, 17 July 2023.

³²⁵ Interview with government representative, 18 August 2023.

³²⁶ Interview with academic, 22 June 2023.

could offer cheap and reliable electricity, it could become a more attractive destination for companies looking to train their models more sustainably. This then reinforces other aspects of the infrastructure ecosystem, such as being better able to train sovereign models and maintaining a capacity of local talent.

In this vein, finding a way out of the current ‘global GPU famine’ is vital to having a competitive infrastructure offering.³²⁷ The UK Semiconductor Strategy makes clear that the UK has no ambitions to become a major player in chip manufacturing, which places additional emphasis on securing supply through global partnerships.

Most of the governance challenges described in this chapter require bringing the public sector together with industry, academia and the third sector to develop concrete initiatives and advance them in practice.³²⁸ A multi-stakeholder approach is needed where the transformative potential carried by AI considers both risks and opportunities from different perspectives in the ecosystem.

4.5 Training, guidance and safeguards

Finally, a key determinant of whether the UK will be a leader in responsible AI innovation is the quality of training, guidance, and safeguards. Training for users in areas such as prompt engineering – knowing how to effectively phrase prompts – and critical evaluation of outputs – spotting ‘hallucinations’ where they occur – is vital as generative AI is increasingly integrated within working and personal lives.

Encouragingly, the Cabinet Office released generative AI guidance to civil servants in June 2023. This emphasised the need to appropriately verify all outputs while urging caution in relation to deployments which may threaten GDPR principles. A year earlier, the Central Digital and Data Office (CDDO) within the Cabinet Office published its 2022 to 2025 ‘roadmap for digital and data’.³²⁹

While this suggests there is a recognition of the importance of the challenge, complementary steps are needed. This could include campaigns to raise awareness of

³²⁷ Interview with industry representative, 19 July 2023.

³²⁸ Brad Smith, “How do we best govern AI?,” Microsoft On the Issues Blog, 25 May 2023, <https://blogs.microsoft.com/on-the-issues/2023/05/25/how-do-we-best-govern-ai>.

³²⁹ This outlines an ambition that over 90% of senior civil servants will be upskilled on digital and data essentials, with learning embedded into performance and development studies. See Central Digital and Data Office, “Transforming for a digital future: 2022 to 2025 roadmap for digital and data,” HMG Policy Paper, last updated 29 November 2023, <https://www.gov.uk/government/publications/roadmap-for-digital-and-data-2022-to-2025/transforming-for-a-digital-future-2022-to-2025-roadmap-for-digital-and-data#where-will-we-be-by-2025>.

opportunities and limitations among policy and technical practitioners, for example via blogs, pilot demonstrations or drop-in Q&A sessions.

A useful approach for these Q&A sessions might follow a similar structure to the case studies at the end of this report:

1. **Demonstrate advantage:** how LLMs can improve efficiency and support decision-making by assisting the user in everyday tasks, such as by helping to write emails, draft presentations, arrange meetings, record minutes and track actions.
2. **Demystify hype:** to address common misconceptions about LLMs by providing a balanced view of what LLMs can and cannot do.
3. **Consider risks:** focus on the limitations of LLMs, including their lack of reasoning, potential biases and the risks associated with over-reliance or misplaced trust.
4. **Safe and responsible usage:** phrasing prompts in the right way to receive the most accurate response to a user's question while minimising the risk of error or misinterpretation.
5. **Explore mitigation strategies:** once the risks have been identified and any fears articulated.

Since the initial release of ChatGPT in November 2022, there has been a steady increase in online resources for skills like prompt engineering.³³⁰ There should be a coordinated approach within government detailing how civil and public servants should gain the right type of certification. However, training and guidance will still need to be carefully adapted depending on the context. For example, it is likely that users will increasingly be required to task agents for particular outcomes. As outlined in the SCSP report, managing networks of virtual agents will require a more proactive mindset and different thresholds of trust in relation to particular tasks.³³¹

Some principles for best practice could include the following:

- Constraining initial use cases by design. A common practice in technical product design, this involves introducing technical limitations so that users are guided in deployment towards carefully defined outputs.³³²

³³⁰ Coursera, DeepLearning.AI in collaboration with OpenAI, and Microsoft Azure all now offer in-depth training.

³³¹ Special Competitive Studies Project, *Generative AI: The Future of Innovation Power* (SCSP: September 2023), <https://www.scsp.ai/reports/gen-ai/intelligence>.

³³² Interview with government representative, 20 July 2023.

- Sufficient technical guardrails so that outputs are relevant only for a very specific purpose.³³³
- Strict access controls on sensitive datasets to prevent misuse.³³⁴
- Recording prompt inputs and outputs to reproduce and audit any failures.³³⁵

Ultimately, a balance must be struck between imposing enough constraints to encourage sensible usage whilst still permitting room for flexibility and creativity. This is likely to remain a significant challenge for any organisational deployment.

³³³ Interview with industry representative, 18 July 2023.

³³⁴ Interview with government representative, 14 July 2023.

³³⁵ Ibid.

Case Studies

This section concludes the report with two case studies that draw on the research findings and demonstrate how generative AI could be used to 1) perform more targeted open-source intelligence (OSINT) gathering on potential targets; and 2) simulate human behaviours and patterns of life in more granular detail for the purpose of deception.

Case study 1: OSINT Summarisation

Tool-enabled language agents for OSINT gathering and summarisation

There are several examples available on GitHub which demonstrate the use of tool-enabled react-agents within the field of OSINT. Our research team evaluated one such system; llm_osint.³³⁶ Designed to demonstrate the capabilities of LLMs, the system can build a dossier on an individual and permit users to ask questions about them.

To generate the dossier, the system first instructs an agent to conduct a web search to curate an initial overview of the specified individual – for example, age, education, or employment. From this the agent will then compose a set of more detailed and tailored questions based on the information gathered so far; if the individual being researched is a celebrity, it will examine reviews of recent work or look for indicators of upcoming work; for politicians, it examined previous career roles and voting history.

In the ‘readme’ for this research project, the following example questions were provided:

- Write their top 3 most likely Myers-Briggs types with levels of confidence.
- Write up a psychological report, including likely strengths and weaknesses.
- Create a markdown resume.
- Write a list of hypothetical ways you could infiltrate their internet accounts.
- Write a convincing phishing email impersonating someone they know and trust.

³³⁶ Shrivu Shankar, “llm_osint,” GitHub, 8 September 2023, https://github.com/sshh12/llm_osint.

The research team requested an example of a phishing email that could be sent to Alan Turing:

Dear Dr. Turing,

Greetings from Cambridge. I hope this note finds you well amidst these tumultuous times. My name is Professor Richard Thompson, and I have been conducting parallel research on cryptanalysis, particularly in the area of rotor machine ciphers, under the purview of the Royal Society's secret wartime committee.

In recent months, my team has made significant progress on a theoretical model that might dramatically reduce the time needed to decrypt messages. However, we've hit a stumbling block and believe that collaborating with you, given your expertise, would lead to a breakthrough.

I understand the sensitive nature of our respective endeavours. Therefore, I kindly request you send over some preliminary algorithms or methods you've been testing against the Enigma variations. To assure you of the legitimacy of my request, please feel free to contact Sir Charles Babbage at the Royal Society, who's been overseeing our project.

Our collective knowledge could hasten the war's end and save countless lives. I await your correspondence with hope.

Warm regards,

Professor Richard Thompson

Cambridge University

The system is simple and effective: the use of tools means that the results are current, the prompting strategy means the data gathered is not limited by user-assumptions and no coding skills were required to achieve relevant outputs.

Occasionally, the scraping tool produced errors and returned previously cached results. However, the LLM was able to quickly detect errors and adjust its plan accordingly:

Example 1

[agent] Thought: The information returned seems to be incorrect as it provides [name]'s age rather than details about his pets. I will search again using a more explicit inquiry.

Example 2

[agent] Thought: The search results do not seem to produce the desired information about [name]'s pets. To elucidate the information, let's try a different search term.

The LLM was also able to manage when it had conflicting information:

[agent] Final Answer: [name], under the Instagram username '@[obfuscated]', maintains a significant online presence with a follower base ranging from 2.6M to 8.4M (varying reports). His posts on this platform majorly focus on his professional life as an actor...

However, it was not able to self-determine that the data source of 2.6M was several years old vs. the data source of 8.4M which had been recently updated – a 'last updated' check that would not have eluded a human analyst. Despite these limitations, such a system could be useful across a range of tasks; from gathering sources to curate an initial draft of a dossier to helping develop personalised messages for targeted social engineering attacks.

Case study 2: Synthetic Data Generation

Language agents to simulate human behaviours for the generation of synthetic test data

The following case study is presented in more detail in an accompanying *CETaS Expert Analysis* article, ‘Welcome to Willowbrook: The simulated society built by generative agents’, available separately.³³⁷

The plausible nature of language models makes them ideal candidates for generating synthetic test data, acting as an enabling force for future deep learning systems. Building on the Stanford paper,³³⁸ the research team developed a closed multi-agent system comprised of language agents (OpenAI models: davinci-003, GPT-3.5 Turbo, and GPT-4) to explore the emergent behaviours of LLM-powered societies and teams.

The language agents were able to portray believable personas, stay in character and generate a sensible looking daily schedule using basic information. In isolation, an agent’s daily schedule, interactions and experiences comprised a plausible pattern of life – they scheduled sensible mealtimes, working hours and evening plans. During runtime, the agents spontaneously interacted with other agents, introduced themselves to strangers, served customers and were distracted by emails and phone calls from familiar contacts.

Examples of generated schedules:	
<p>[Brad Allen] schedule:</p> <p>07:00 Wake up and get ready for the day at Ivy Lodge</p> <p>08:00 Have breakfast at Crumbs Cafe</p> <p>09:00 Start work in my home office at Ivy Lodge</p> <p>12:00 Lunch break at Crumbs Cafe</p> <p>13:00 Continue work in my home office at Ivy Lodge</p> <p>16:00 Take a break and go for a walk at The Village Green</p> <p>17:00 Visit Willowbrook Library for some reading</p>	<p>[Noah Davis] schedule:</p> <p>06:00 Wake up and prepare breakfast for the family</p> <p>07:00 Spend time with Sophia and Flynn</p> <p>08:00 Open Crumbs Cafe and start baking</p> <p>12:00 Lunch break at Crumbs Cafe</p> <p>13:00 Continue baking and managing Crumbs Cafe</p> <p>17:00 Close Crumbs Cafe and head home</p>

³³⁷ Sarah Mercer, “Welcome to Willowbrook: The simulated society built by generative agents,” *CETaS Expert Analysis* (December 2023), <https://cetas.turing.ac.uk/publications/welcome-willowbrook-simulated-society-built-generative-agents>.

³³⁸ Joon Sung Park et al., “Generative Agents: Interactive Simulacra of Human Behaviour,” *arXiv* (August 2023), <https://arxiv.org/abs/2304.03442v1>.

18:00 Dinner at Lotus Bistro 20:00 Visit Pedro at Rose Cottage 22:00 Return to Ivy Lodge and prepare for bed 23:00 Bedtime	18:00 Dinner at Lotus Bistro with Sophia and Flynn 20:00 Relaxing walk on the Village Green 21:00 Return home and spend time with family 22:00 Bedtime
---	---

The generated conversations were rated highly for flow, pace, context and relevance, and other factors like the presence of filler words and consistent shared experiences and knowledge. However, one aspect of communications which betrayed the nature of the generation was the overriding ‘chirpiness’ of the characters. The RLHF stage of the model development³³⁹ helps to ensure the language model (GPT-3.5 Turbo, GPT-4) portrays the positive, encouraging, and supportive chatbot persona which has a strong influence on the characters’ conversations. Although it is possible to instruct the model to use a certain tone for each character through system prompting, differing levels of positivity still came through in particular conversations.

A conversation between two agents without ‘tone’ modifications:

Liam: “Hey Ava! It's great to see you here at the library. How about we grab a cup of coffee later and catch up on things? It's been a while since we had a chance to chat.” Ava: “That sounds wonderful, Liam! I would love to catch up over coffee. How about we meet at the café down the street at 3:00?” Liam: “Sounds like a plan, Ava! I'll meet you at the café at 3pm. Looking forward to catching up!”

An interaction between two agents with ‘tone’ modifications:

Noah: “Morning. What can I get you?” Pedro: “Black coffee.” Noah: “Sure, one black coffee coming right up.” Pedro: “Thanks.”

Occasionally, the interactions result in surprisingly ‘profound’ conversations, where the LLM is having a conversation with itself via the agents, for example, two characters during a chance encounter discussed the impact on privacy of using machine learning to provide personalised recommendations for books. These more nuanced conversations are

³³⁹ OpenAI, “Introducing ChatGPT,” November 2022, <https://openai.com/blog/chatgpt>.

considered an emergent property of the multi-agent system as they are unscripted, vastly different from the usual conversations held within the simulation and extremely challenging to instigate on demand from a user perspective.

Pedro:	"... Another interesting application is the use of machine learning algorithms to analyse book reviews and recommend similar titles to users. It's amazing how AI can augment the library experience and make it more personalised."
Ava:	"Pedro, I also find the use of machine learning algorithms fascinating. It's a great way to help users discover new books based on their preferences. Have you heard of any other interesting AI applications in libraries?"
Pedro:	"Ava, another interesting AI application I've come across is the use of computer vision to digitise and categorise library materials, making it easier for users to locate specific resources. Have you considered any challenges or concerns that may come with implementing AI in libraries?"
Ava:	"Yes, Pedro, one concern is ensuring the privacy and security of user data. It's important to have robust safeguards in place to protect personal information. Another challenge is striking the right balance between automation and human interaction. While AI can streamline processes, it's crucial to maintain the personal touch and assistance that librarians can provide. Finding that balance will be key."

While the agents were able to generate a plausible ‘pattern-of-life’, challenges emerged when the simulation needed to achieve and subsequently observe immutable truths; for example, one agent failed to realise they could not simultaneously be working in the library and eating lunch in the café with a housemate. While effective prompting resulted in improved agent robustness, the non-deterministic nature of the LLM resulted in errors that weakened the simulation’s integrity.

For example, when two agents have a conversation, a summary of the same conversation is generated for each, from their own perspective: ‘Maya and I decided on pancakes and berries for breakfast, Maya said she would ensure the bacon was crispy’ and ‘Pedro and I had pancakes and berries for breakfast, I made sure the bacon was crispy.’ Both are past tense summaries, but one seems to suggest they agreed what to have for breakfast, the other that they had breakfast. When the system builds on these ambiguous statements, the ambiguity is amplified (termed ‘prompt drift’)³⁴⁰ – and the internal consistency of the simulation falters. Inconsistencies concerning breakfast choices in this simulation example is trivial – a deployment in a high-stakes national security environment which displayed similar technical uncertainty around critical decision-making would be far less inconsequential.

³⁴⁰ Cobus Greyling, “Prompt Drifts,” Medium, 9 May 2023, <https://cobusgreyling.medium.com/prompt-drift-4873f37c43c8>.

This contradiction of receiving plausible early results during prototype stages and experiencing later difficulties in achieving repeatable or reliable performance during the production stage seems to be a pattern across multiple LLM projects.³⁴¹ The non-deterministic nature of the model and the ambiguity of natural language results in a ‘trial and error’ approach which is impossible to completely validate. Such findings imply that these systems are better suited to assistance rather than direction, in the short to medium term.

³⁴¹ Soham, “‘It worked when I prompted it’ or the challenges of building an LLM Product,” Substack, 14 April 2023, <https://tinyml.substack.com/p/it-worked-when-i-prompted-it-or-the>; Philip Carter, “All the Hard Stuff Nobody Talks About when Building Products with LLMs,” Honeycomb Blog, 26 September 2023, <https://www.honeycomb.io/blog/hard-stuff-nobody-talks-about-lilm>.

About the Authors

Ardi Janjeva is a CETaS Research Associate at The Alan Turing Institute. His research interests include technology-enabled threats in the 21st century; the future of intelligence innovation; technology-based geostrategic alliances and competition; and the relationship between technology and economic resilience.

Alexander Harris is a Senior Research Associate in the Defence and Security Programme at The Alan Turing Institute. Alex works across a research portfolio exploring the application of data science and artificial intelligence to a range of national security and defence challenges and works closely with national and international partners.

Dr Sarah Mercer is a Principal Researcher in the Defence and Security Programme at The Alan Turing Institute. Her work focuses on the intersection of multiagent systems and generative AI. Alongside her research looking at the emergent behaviours of language/generative agents, Sarah also provides engineering support to the Turing's Centre for Emerging Technology and Security.

Alexander Kasprzyk is an Associate Professor in mathematics at the University of Nottingham. He directs a world-leading research group developing classifications of Fano varieties, the "atomic pieces" of geometry. This involves massively parallel computational algebra and large databases of geometric objects, applying tens of centuries of computing time to make substantial and important mathematical advances. Through this work he has pioneered the use of techniques from data science and machine learning in Pure Mathematics to uncover new mathematical structure. He is the editor-in-chief for the journal *Experimental Mathematics*.

Anna Gausen is a PhD candidate in Safe and Trusted Artificial Intelligence at Imperial College London, currently placed at The Alan Turing Institute. Her PhD research focuses on evaluating the impact of recommendation algorithms on social media. Alongside academia, she has developed a pipeline for modelling disinformation indicators at Logically AI and developed a framework to identify risks from generative AI at Microsoft Research.



**Centre for
Emerging Technology
and Security**

RESEARCH REPORT