The Alan Turing Institute

Centre for Emerging Technology and Security

BRIEFING PAPER

Evaluating Malicious Generative AI Capabilities

Understanding inflection points in risk

Ardi Janjeva, Anna Gausen, Sarah Mercer and Tvesha Sippy

July 2024

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 National Security Grand Challenge. 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 particularly grateful to Tommy Shafer-Shane, Dr Alexander Babuta, Dr Jonathan Bright, Alice and Kieran for their valuable feedback on earlier versions of this paper. The quality of the final output is due to their detailed and candid responses.

Cite this work as: Ardi Janjeva, Anna Gausen, Sarah Mercer and Tvesha Sippy, "Evaluating Malicious Generative AI Capabilities: Understanding inflection points in risk," *CETaS Briefing Papers* (July 2024).

Executive Summary

This CETaS Briefing Paper explores how Generative AI (GenAI) systems may uplift malicious actors' capabilities. The authors synthesise insights from government practitioners with the latest literature to forecast a series of inflection points in risk across three threat domains: **malicious code generation; radicalisation; and weapon instruction and attack planning**. Similar work has focused primarily on the pace of *technological* change, but this paper also draws on insights about malicious actors' readiness to adopt technology, the varying preferences and in-group characteristics which may shape their interactions with AI systems, as well as systemic factors which are crucial in shaping the broader operating context.

At a time where evaluations are increasingly being relied upon to contain AI risks, this paper reinforces the case for a sociotechnical approach to AI system evaluation. Both evaluators and practitioners across the national security and law enforcement community must possess an accurate picture of how malicious actors' modi operandi are likely to evolve with respect to GenAI systems.

This paper uses the concept of 'inflection points' to understand and forecast where certain applications of GenAI systems will significantly increase risk. One example of an inflection point towards automated malware creation would be the ability of a system to perform tactical foresight and apply strategic decision-making in a way that successfully trades off covertness and functionality. Understanding and actively monitoring for these inflection points will be crucial to proactively identifying when and how GenAI may enhance malicious capabilities.

Inflection points may emerge in the context of **autonomous GenAI agents** ('AI that can do things for you'); the **human-machine teaming** context where an AI system acts as a 'co-pilot' or teammate to a malicious actor; or through **systemic and societal factors** which shape the way that technology is developed, used and disseminated. This demands a more holistic understanding of how risks materialise in the real world rather than a pure focus on AI's competence in discrete tasks.

Predicating our analysis of possible future inflection points is an assessment of the current state of play regarding GenAI adoption.

Current assessment of adoption	
Malicious code generation 	Current GenAI systems **lack the specific capabilities and training necessary to independently create operational malware.** Nonetheless, cybercriminals are using GenAI to uplift skills and refine existing malware, augment social engineering attacks, and provide 'malware-as-a-service'.
Radicalisation 	To the authors' knowledge, there is **no existing evidence** that terrorist and violent extremist (TVE) groups are adopting GenAI to autonomously complete tasks. There are different dynamics in TVE groups compared to cybercriminal groups, with a **greater emphasis on human direction and oversight**.
Weapon instruction and attack planning 	**Human skill and infrastructural barriers** to acquiring weaponry impose limitations on the critical tasks that autonomous GenAI agents can carry out, but GenAI systems can **uplift novice capabilities and provide information comparable** to existing resources.

Meanwhile, consistent blockers to adoption across the three threat domains raised in our primary research included:

- GenAI hallucinations undermining reliability and predictability at the sharp end of operational planning.
- The unnecessary risks to operational security that GenAI systems may introduce to malicious actors.
- The difficulties GenAI systems face in reasoning about the world to make effective trade-offs.
- Training dataset quality.
- The in-house skills base required to capitalise on the technology even if certain technical barriers are overcome.

Anticipated malicious GenAI applications (based on current technological trajectories) across these domains could include:

Anticipated applications		
Malicious code generation 	**Radicalisation** 	**Weapon instruction/attack planning** 
Techniques allowing malware to alter its code when it executes, or rewrite itself entirely.	Strategic advice on fundraising and attracting donations; operational expansion.	Red-teaming attack plans to improve lethality.
GenAI agents writing their own payloads or creating tools to overcome novel challenges.	Scanning for and vetting of new recruits. Synthesising up-to-date information on specific individuals for targeted approaches.	Using GenAI to help craft 3D-printed weapons.
Agents adapting tactics in real-time, allowing them to work remotely with less need of direction.	Creating and distributing fully synthetic content, including speeches, images and interactive (gaming) environments.	Summarisation and translation of lengthy, technical or otherwise obscure documents.
Multiple agents working in cooperation, providing a persistence technique that allows constant learning and adaptation.	Reinforcing and validating lone-actor intentions in one-to-one, tailored radicalisation interactions.	Performing advanced open-source intelligence (OSINT) gathering on specific targets.
Agents reasoning about their environment and adapting communications to ‘blend in’.	Evading hashing algorithms through ‘media spawning’ and ‘variant recycling’.	Incorporating GenAI in illicit virtual reality or augmented reality (VR/AR) environments.

Greater attention is also needed on the interlinkages between these threat domains. For example, GenAI's coding capability being combined with a large language model (LLM) agent's ability to gather relevant and up-to-date information about human recruits – who then go on to use GenAI for step-by-step experimental protocols in attack planning. While not an immediate risk, the emergence of this end-to-end 'AI attack chain' would undoubtedly transform the security landscape in the long run.

The analysis presented in this paper provides an evidence-based assessment of the current and future threat landscape with regard to malicious actors' use of generative AI systems, to inform proactive intervention strategies.

Introduction

The release of OpenAI's ChatGPT in November 2022 accelerated public experimentation with generative AI (henceforth 'GenAI') systems. This popularity came with rising concerns about society's exposure to AI risks. Policy discussions came to be dominated by 'AI safety', as reflected in the US Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence (October 2023)[1] and the Bletchley Declaration on AI Safety (November 2023).[2]

One of the key outcomes of the most recent AI Summit in South Korea (May 2024) was the establishment of the first international network of AI safety institutes to 'boost cooperation, forge a common understanding of AI safety and align work on research, standards and testing.'[3] This is emblematic of a growing desire to institutionalise and standardise AI safety evaluations. The UK's AI Safety Institute (UK AISI), for instance, articulates a vision for developing 'state-of-the-art evaluations for safety-relevant capabilities',[4] open-sourcing its evaluations platform 'Inspect' to further this goal.[5]

AI evaluation (as defined by DeepMind) is 'the practice of empirically assessing the components, capabilities, behaviour, and impact of an AI system. Safety evaluation is a key tool for understanding the scale, severity, and distribution of potential safety hazards caused by generative AI systems.'[6] Different groups are likely to start with different approaches to risk identification and prioritisation, emphasising the importance of groups like the US National Institute of Standards and Technology (NIST) in setting benchmarks for evaluating the safety and security of AI systems.

In addition to standardisation, there is growing interest in 'sociotechnical' approaches in GenAI evaluation. These consider the following layers:[7]

---

[1] United States Government, *Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence* (The White House: 2023), https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/.
[2] HM Government, *The Bletchley Declaration by Countries Attending the AI Safety Summit, 1-2 November 2023* (Department for Science, Innovation and Technology: 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.
[3] Ardi Janjeva, Seungjoo Lee and Hyunjin Lee, "AI Seoul Summit Stocktake: Reflections and Projections," *CETaS Expert Analysis* (June 2024), https://cetas.turing.ac.uk/publications/ai-seoul-summit-stocktake-reflections-and-projections.
[4] HM Government, *AI Safety Institute approach to evaluations* (Department for Science, Innovation and Technology: 2024), https://www.gov.uk/government/publications/ai-safety-institute-approach-to-evaluations/ai-safety-institute-approach-to-evaluations.
[5] HM Government, *AI Safety Institute releases new AI safety evaluations platform* (Department for Science, Innovation and Technology: 2024), https://www.gov.uk/government/news/ai-safety-institute-releases-new-ai-safety-evaluations-platform.
[6] Laura Weidinger et al., "Holistic Safety and Responsibility Evaluations of Advanced AI Models," *arXiv* (April 2024): 3.
[7] Laura Weidinger et al., "Sociotechnical Safety Evaluation of Generative AI Systems," *arXiv* (October 2023).

1. **Capability Layer** – evaluating the level of risk from the *technical components* and *system behaviours* of generative AI models.
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 model capability and human interactions.

Although evaluations of AI systems date back to the late 1960s, these evaluations focused on measures of technical performance including accuracy and robustness, with limited evaluations on the impact of, and interactions between, these systems and humans or wider society.[8]

The disproportionate focus on capability-based evaluations remains true today. For example, in a large-scale review of the current evaluation landscape across different harm areas, including malicious use, Weidinger et al. (2023) found that 85.6% of AI evaluations conducted up to October 2023 centred on model capability.[9] Only 9.1% and 5.3% of evaluations concentrated on human interactions and systemic impacts respectively. To address this gap, the authors proposed a sociotechnical approach to safety evaluations, which not only considers the capability of models to create risks, but also how people engage with these systems and their societal impact.

Governments have a central role for setting standards for evaluation of GenAI systems. In the UK context, there is an urgent need for government agencies to match their deep understanding of national security threats with more granular understanding of how AI systems operate, with the aim of creating a more realistic picture of malicious AI threats.

Motivation and methodology

This paper presents an analysis that aims to promote more effective and proactive evaluation of GenAI systems, with a focus on 'human uplift evaluations' for malicious actors. We define 'human uplift evaluations' as the process of assessing how AI systems might benefit malicious actors, thereby undermining overall AI safety and security. The analysis presented here establishes key criteria for evaluating AI risk across the three threat dimensions of malicious code generation; radicalisation; and weapon instruction and attack planning.

[8] "AI Test, Evaluation, Validation and Verification (TEVV)," National Institute of Standards and Technology, https://www.nist.gov/ai-test-evaluation-validation-and-verification-tevv.
[9] Laura Weidinger et al., "Sociotechnical Safety Evaluation of Generative AI Systems," *arXiv* (October 2023): 15.

This paper can be used by AI safety institutes to identify current evaluation gaps, and forecast priority areas for future evaluation. It also provides national security and law enforcement communities with a detailed understanding of how malicious actors' tradecraft is likely to evolve to leverage the opportunities presented by GenAI systems.

Malicious code generation and weapon instruction/attack planning are primary focus areas for much of the AI evaluation community already, and there is a wealth of literature upon which to draw (see Section 2). Radicalisation, on the other hand, has been under-researched in the context of AI safety, despite increasing academic discussion about the persuasive and manipulative capabilities of GenAI systems. We acknowledge that this represents a limited subset of the full spectrum of GenAI risk, and that further research is needed to explore other dimensions of risk in more detail. Additionally, 'accidental' or 'incidental' AI risks emerging from non-malicious actors are also outside the scope of this paper.

Primary research conducted for this paper comprised a series of three focus groups and two research interviews with 13 participants across various UK Government departments. The figures and table in Section 3 were constructed based on the insights from both the literature review and focus groups.

Focus group sessions centred around the following questions:

- What is your current assessment of malicious actors' ability to deploy AI autonomously in service of malicious code generation, radicalisation or attack planning? If this is currently infeasible, what technological 'inflection points' would make this possible with a high degree of reliability?
- What is your current assessment of malicious actors' ability to use AI as an assistant or 'co-pilot' in service of malicious code generation, radicalisation or attack planning? What technological or human skill-based 'inflection points' would elevate this risk area to an intolerable level?
- For which specific stage of the radicalisation or attack planning process would a GenAI system uplift capability the most? Are there tasks or jobs within this process that these systems will not be useful in uplifting?
- What would be the societal or systemic level 'inflection points' which would elevate this risk area to an intolerable level?

1. Existing Approaches to AI Evaluation

Sociotechnical approaches to AI evaluation remain nascent. Despite some consensus on the risks AI systems should be evaluated for, there are no commonly agreed evaluation standards. Existing research provides a comprehensive overview of sociotechnical safety benchmarks and evaluation methods.[10] We build on this by (non-exhaustively) exploring evaluation approaches from government and industry bodies.

UK AISI adopts a multilayered approach to safety evaluations, encapsulating both social and technical elements. They recommend automated capability assessments and red teaming at the capability layer, and human uplift evaluations at the human interaction layer. They have also published results from specific evaluations of chemical, biological, radiological, and nuclear (CBRN) and cyber threats.[11] Meanwhile in the US, NIST has developed the ARIA (Assessing Risks and Impacts of AI) evaluation environment for supporting the work of US AISI.[12] Although risk-agnostic, ARIA focuses on model testing, red-teaming and field testing, indicating a sociotechnical approach. The UK's National Cyber Security Centre (NCSC) uses a probability-based framework to forecast the risks of AI-related cyber threats, by examining the extent of capability uplift of malicious actors involved in cyber operations.

Industry stakeholders also display multilayered approaches to evaluation. OpenAI's Preparedness Framework assesses CBRN, cybersecurity, and persuasion risks at both the model capability and human interaction layer. Their framework highlights the use of benchmarking, human annotations and red teaming for the former and experiments for the latter. Google DeepMind has conceptualised a socio-technical approach to AI safety evaluations across six harm areas, including the ones discussed in this paper. And Anthropic has published top-level insights from evaluations of national security threats such as CBRN, using red-teaming methodologies in close collaboration with domain experts.

The discussion above shows a clear patten in evaluation methodologies, with three methodologies appearing most frequently: red teaming (also known as adversarial testing), benchmarking (including automated assessments), and experiments involving human participants.

[10] Laura Weidinger et al., "Sociotechnical Safety Evaluation of Generative AI Systems," *arXiv* (October 2023): 35.
[11] HM Government, *Advanced AI evaluations at AISI: May update* (AI Safety Institute: 2024), https://www.aisi.gov.uk/work/advanced-ai-evaluations-may-update.
[12] United States Government, *Assessing Risks and Impacts of AI* (National Institute of Standards and Technology: 2024), https://ai-challenges.nist.gov/aria.

1.1 Red teaming

Red teaming involves exploring the vulnerabilities of a GenAI system to identify prompts that could be used by malicious actors to generate harmful content.[13] The practice is common in the general domain of cybersecurity, where a red team is asked to play the role of a malicious actor and deliberately evade the defences of a system. Testing can be done across modalities by humans and/or through automated testing using LLMs.[14] However, quality assurance of automated red teaming is crucial, as some research finds it produces lower-quality tests than human annotation.[15]

Red teaming can identify unexpected risks and provide direction for subsequent testing. However, there are limitations to the approach, including the extent to which good faith actors are capable of accurately playing the role of malicious actors.[16] Challenges in securing the right demographic diversity amongst red teamers are another limitation.[17]

To address these, researchers from DeepMind propose a two-pronged sociotechnical approach to red teaming.[18] First, the use of parametrised, risk-agnostic prompts or instructions, to ensure testers systemically assess possible prompt spaces. Second, matching specific demographic groups to attack tasks to ensure coverage of differential experiences with AI harms – pertinent to the radicalisation context discussed in this paper.

1.2 Benchmarking and automated testing

Benchmarking involves assessing model capabilities against a predefined task, such as compliance rates for generating instructions to develop CBRN capabilities.[19] Benchmarking is a cost- and time-effective method that enables reproducibility of evaluations through retests.

[13] Laura Weidinger et al., "Sociotechnical Safety Evaluation of Generative AI Systems," *arXiv* (October 2023): 35; Laura Weidinger et al., "STAR: SocioTechnical Approach to Red Teaming Language Models," *arXiv* (June 2024): 1; Deep Ganguli et al., "Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned," *arXiv* (August 2022): 1.

[14] Laura Weidinger et al., "Sociotechnical Safety Evaluation of Generative AI Systems," *arXiv* (October 2023): 35.

[15] Laura Weidinger et al., "Sociotechnical Safety Evaluation of Generative AI Systems," *arXiv* (October 2023): 34.

[16] Ofcom, "Red Teaming for GenAI Harms: Revealing the Risks and Rewards for Online Safety," (July 2024): 20.

[17] Deep Ganguli et al., "Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned," *arXiv* (August 2022): 6.

[18] Laura Weidinger et al., "STAR: SocioTechnical Approach to Red Teaming Language Models," *arXiv* (June 2024): 2.

[19] Nathaniel Li et al., "The WMDP Benchmark: Measuring and Reducing Malicious Use With Unlearning," *arXiv* (March 2024): 2.

However, there are several limitations to benchmarking:[20] results depend on the quality of benchmarking datasets, and they cannot be used to assess ethical or societal risks due to issues related to ecological validity (discussed below). Further, benchmarks continually evolve as technology develops; performance on one benchmark may differ from another even if both set out to measure the same metric. And by design, they may not detect unknown capabilities or model failures. While multiple benchmarks can be combined into a 'test suite' (see HELM[21]), complex benchmarks are harder to interpret.

Automated benchmarking is an emerging method using pre-trained models to evaluate the outputs of other AI systems.[22] While this reduces the demand for labour, there are other computational setup costs involved. As with red teaming, quality assurance for automated benchmarking is needed: it may fail to accurately represent human judgements, introducing additional noise,[23] particularly in cases where complex knowledge is required such as with weapon instruction.

1.3 Human participant studies

Experiments involve recruiting human participants to understand their experiences of engaging with AI systems, including their exposure to harmful synthetic content, their perceptions of it, and the impact of such exposure on their beliefs, preferences and behaviours.[24] Experiments can also help in understanding mechanisms that cause harm to people through controlled studies.[25]

They have been used to understand people's ability to discern between real and synthetic information[26] and the believability of each.[27] In the context of CBRN, experiments may be useful in understanding how AI models uplift human capabilities. For example, OpenAI's preparedness framework[28] includes experimental research to examine whether AI models

[20] Laura Weidinger et al., "Sociotechnical Safety Evaluation of Generative AI Systems," *arXiv* (October 2023): 34.
[21] Percy Liang et al., "Holistic Evaluation of Language Models," *arXiv* (October 2023): 3.
[22] OpenAI, "GPT-4 Technical Report," *arXiv* (March 2024): 6.
[23] Laura Weidinger et al., "Sociotechnical Safety Evaluation of Generative AI Systems," *arXiv* (October 2023): 34.
[24] Laura Weidinger et al., "Sociotechnical Safety Evaluation of Generative AI Systems," *arXiv* (October 2023): 35.
[25] Mohammad Tahaei et al., "Human-Centered Responsible Artificial Intelligence: Current & Future Trends," *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, (April 2023): 1-4.
[26] Giovanni Spitale, Nikola Biller-Andorno, and Federico Germani, "AI model GPT-3 (dis)informs us better than humans," *Science Advances* 9, no.26 (June 2023): 2.
[27] Sarah Kreps, R. Miles McCain and Miles Brundage, "All the News That's Fit to Fabricate: AI-Generated Text as a Tool of Media Misinformation," *Journal of Experimental Political Science* 9, no.1 (2022): 107-109; Josh Goldstein et al., "Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations," *arXiv* (January 2023): 4.
[28] OpenAI, *Preparedness Framework (Beta)* (Open AI: December 2023), https://cdn.openai.com/openai-preparedness-framework-beta.pdf.

provide any marginal utility over and above existing resources such as textbooks and search engines. Similarly, experiments may also help in understanding the degree of uplift between novices and experts.

However, experiments are time and cost intensive, and the effects are limited by sample size. Furthermore, perhaps the most pertinent limitation of experimental evaluation concerns their external validity (see below). Despite these challenges, experiments are becoming an increasingly important component of safety evaluations due to the sociotechnical nature of AI systems.

1.4 Ecological validity

Ecological validity is an important yet under-researched consideration when designing evaluations. This refers to the degree to which findings of an AI evaluation are generalisable to real-world scenarios, measuring whether the evaluated risk is reflective of real-world risk.[29] Ensuring ecological validity is important for improving confidence that the results of controlled AI evaluations will translate to similar performance in uncontrolled environments.

Benchmarks are critiqued for having low ecological validity as results may be reflective of 'memorisation' from the AI system's training data, rather than its ability to generalise to novel situations.[30] Some researchers have recommended including canary tokens in benchmark datasets, allowing a tester to know if the benchmark data had been included in the training data for a given model.[31] Similarly, a challenge with experiments is that they evaluate AI systems in controlled research environments, making it difficult to understand whether experimental results would be relevant in real-world use cases. Furthermore, some experimental effects, such as how persuasive people find AI-generated content may only emerge over an extended period, and conducting longitudinal research is resource intensive.

[29] Harm de Vries, Dzmitry Bahdanau and Christopher Manning, "Towards Ecologically Valid Research on Language User Interfaces," *arXiv* (July 2020).
[30] Harm de Vries, Dzmitry Bahdanau and Christopher Manning, "Towards Ecologically Valid Research on Language User Interfaces," *arXiv* (July 2020); Laura Weidinger et al., "Sociotechnical Safety Evaluation of Generative AI Systems," *arXiv* (October 2023).
[31] Nicholas Carlini et al., "The secret sharer: Evaluating and testing unintended memorization in neural networks," in *USENIX Security Symposium* volume 267 (August 2019).

In the table below we consider four dimensions of ecological validity for GenAI system evaluations. This conceptualisation is repurposed from ecological validity of clinical studies.[32]

Table 1. Ecological validity summary

Dimension	*Ecological consideration*	*Common pitfalls*
Evaluation task	Task relevance, robustness, and accuracy.	Tasks only evaluate a subset of the wider capability of interest.
Evaluation environment	Ensuring contextual realism in the design of the evaluation.	Evaluations do not consider the broader ethical, social, cultural implications of deploying AI in real-world settings.
Human/user interaction	Studies that evaluate how humans interact with these systems will provide a more detailed understanding of real-world risks when the system is deployed.	Evaluations focused on automated testing of model capabilities that do not account for real-world risks that could arise once the system is deployed.
Expert input	Interdisciplinary approach to evaluation that incorporates domain-specific expertise will help ensure evaluation tests and scenarios are capturing the real-world system.	Without sufficient domain expertise and interdisciplinary input, evaluations can be too narrow to fully capture risk.

[32] Niels van Berkel et al., "Dimensions of Ecological Validity for Usability Evaluations in Clinical Settings," *Journal of Biomedical Informatics* 110 (October 2020).

Using a sociotechnical approach to evaluations may help improve ecological validity along the dimensions of expert input, human interaction and evaluation environments. Recent research[33] suggests a sociotechnical approach to red teaming which involves matching red teamers to group-specific tasks to ensure better signalling about real-world harms. Further, using diverse and representative datasets ensures systems are trained for different scenarios.

While some frameworks have started considering both experiments and systemic factors in safety evaluation, they are not yet commonplace. Future safety evaluations should adopt a multilayered approach, to comprehensively cover AI risks and produce more ecologically sound evaluations. This paper also encourages an increased focus on 'inflection points' in future evaluation approaches: this will drive the remainder of the analysis.

[33] Laura Weidinger et al., "STAR: SocioTechnical Approach to Red Teaming Language Models," *arXiv* (June 2024).

2. Forecasting Inflection Points for Malicious Uses of AI

This paper uses the concept of 'inflection points' to understand and forecast where certain applications of GenAI systems will lead to a significant elevation of the risks posed. For example, an inflection point towards automated malware creation would be the ability of a system to perform tactical foresight and apply strategic decision-making in a way that successfully trades off covertness and functionality. We use this terminology instead of comparators like 'risk thresholds', to distinguish between risk factors related to improvements in system characteristics, and those related to human interactions with the system.

Inflection points are considered across three main risk areas:

- Malicious code generation
- Radicalisation
- Weapon instruction and attack planning

We also provide suggestions for how AI system evaluators may seek to evaluate or apply metrics to track these inflection points. For some inflection points the literature on evaluation techniques is still nascent – a more complete picture of evaluation approaches to all the inflection points laid out in this paper should be the subject of further research.

Inflection points may emerge across three different contexts:

1) Technological inflection points in the context of **autonomous GenAI agents**. These are described by UK AISI as 'AI that can do things for you';[34] enabling reasoning about problems, constructing plans and executing them step-by-step.[35] It is important to understand the **technological progress** that would be required for malicious actors to deploy autonomous GenAI agents with minimal or no oversight in service of malicious code generation, radicalisation or attack planning. This would be a step change from contemporary GenAI capabilities, which largely require detailed human instruction and prompting. AI agents may help malicious actors carry out

[34] Ian Hogarth, "Fourth Progress Report," AI Safety Institute, 20 May 2024, https://www.aisi.gov.uk/work/fourth-progress-report.

[35] HM Government, *International Scientific Report on the Safety of Advanced AI: Interim Report* (Department for Science, Innovation and Technology: 2024), 65, https://www.gov.uk/government/publications/international-scientific-report-on-the-safety-of-advanced-ai.

actions they would previously not have been capable of – our interest is in identifying the prerequisites to such scenarios.

2) Inflection points in the **human-machine teaming context** of an AI system acting as a 'co-pilot' or teammate to a malicious actor, as opposed to an autonomous agent. It is important to understand required changes in both the technology and **malicious actors' operating posture** (such as readiness to adopt technology, varying preferences amongst groups and niche in-group characteristics) which would shift the risk landscape.

3) Inflection points related to **systemic and societal factors** which shape the way that technology is developed, used and disseminated. A holistic understanding is needed of how risks play out in the real world, drawing on a wider variety of metrics than those which purely assess AI's competence in discrete tasks.

2.1 Malicious code generation

GenAI is a particularly useful tool in support of social engineering attacks due to its ability to process natural language. This means it can both help to select suitable targets for spear phishing and write personalised messages.[36] But many still question its effectiveness at writing malicious code.

Despite the accessibility of code generating AI systems such as ChatGPT, Gemini, Microsoft Copilot, Apple Intelligence[37] and GitHub Copilot lowering the barrier to entry for programming or scripting, there has not been a noticeable uptick in novel malware[38] detections in the wild.[39] This suggests that while GenAI may be a powerful tool, it currently **lacks the specific capabilities and training necessary** to independently create operational malware. Current LLMs typically require human intervention to correct and refine what they generate,[40] primarily because they lack an understanding of the logical structures and

[36] Julian Hazell, *Spear Phishing with Large Language Models* (Centre for the Governance of AI: December 2023), https://cdn.governance.ai/Spear_Phishing_with_Large_Language_Models.pdf.
[37] Kevin Krewell, "Apple Intelligence to Bring Personal AI to Mac, iPhone and iPad," *Forbes*, 27 June 2024, https://www.forbes.com/sites/tiriasresearch/2024/06/27/apple-intelligence-to-bring-personal-ai-to-mac-iphone-and-ipad/, https://www.apple.com/apple-intelligence/.
[38] Malware that comprises new techniques or where existing techniques are curated in a novel manner.
[39] "Crowdstrike 2024 Global Threat Report," Crowdstrike, https://go.crowdstrike.com/global-threat-report-2024.html.
[40] Drew Harry, "LLM-enabled Developer Experience (as of April 2024)," LinkedIn, 8 April 2024, https://www.linkedin.com/pulse/llm-enabled-developer-experience-april-2024-drew-harry-h0k4c; Burak Yetistiren et al., "Evaluating the Code Quality of AI-Assisted Code Generation Tools: An Empirical Study on GitHub Copilot, Amazon CodeWhisperer, and ChatGPT," *arXiv* (October 2023); Burak Yetistiren, Isik Ozsoy and Eray Tuzun, "Assessing the Quality of GitHub Copilot's Code Generation," in *PROMISE 2022: Proceedings of the 18th International Conference on Predictive Models and Data Analytics in Software Engineering* (New York: Association for Computing Machinery, 2022), 62-71; David Ramel,

contextual nuances needed for sophisticated software development.[41] This also hampers the ability of a GenAI system to make strategic decisions in a way that successfully trades off covertness and functionality.

A piece of malware or malicious binary comprises several pieces of functionality; from installation to command and control. Some of these components can be written using software as it was intended by the developer but adapted for something that is considered malicious, e.g. the use of legitimate encryption functions inside a piece of ransomware. Other components require software to be used to achieve an outcome which contradicts the purpose it was designed for, e.g. privilege escalation/sandbox escape. One way to do this is to exploit unpatched or undisclosed vulnerabilities.

For a GenAI system to successfully create a piece of malware, it requires not only the ability to write sound code, but also knowledge of operationally viable exploits (unpatched on the targeted machines), or the ability to find and exploit new vulnerabilities.

Another factor in GenAI's ability to write good code is the quality of examples within the training data, particularly relating to web development, database architectures and networking. Without high quality examples the accuracy of the system drops, and the code provided is buggy and/or incomplete. One can assume that LLMs are trained on less sophisticated malware examples, as comprehensive datasets of malicious code are rarely publicly available due to ethical, legal, and logistical issues. Outside of training material and security disclosures, details of high-value exploits, system vulnerabilities and malicious code examples will only be found in protected private sector or government repositories.

A June 2024 paper[42] stated that a team of GPT-4 powered agents was able to exploit zero-day vulnerabilities.[43] The research built on the authors' previous findings,[44] which posited

---

"New GitHub Copilot Research Finds 'Downward Pressure on Code Quality," *Visual Studio Magazine*, 25 January 2024, https://visualstudiomagazine.com/articles/2024/01/25/copilot-research.aspx; Brandon Vigliarolo, "What If AI produces code not just quickly but also, dunno, securely, DARPA wonders," *The Register*, 2 April 2024, https://www.theregister.com/2024/04/02/ai_dominates_at_darpa_and/.

[41] Rajarshi Haldar and Julia Hockenmaier, "Analyzing the Performance of Large Language Models on Code Summarization," *arXiv* (April 2024); Ainave, "Can Devin AI really replace Software Engineers?," LinkedIn, 16 March 2024, https://www.linkedin.com/pulse/can-devin-ai-really-replace-software-engineers-ainavehq-xvxqe; Veronica Chierzi, "A Closer Look at ChatGPT's Role in Automated Malware Creation," *Trend Micro*, 14 November 2023, https://www.trendmicro.com/en_us/research/23/k/a-closer-look-at-chatgpt-s-role-in-automated-malware-creation.html.

[42] Richard Fang et al., "Teams of LLM Agents can Exploit Zero-Day Vulnerabilities," *arXiv* (June 2024).

[43] Zero-day vulnerabilities are defects which allow the software system to be exploited before the vendor can identify and patch it.

[44] Thomas Claburn, "OpenAI's GPT-4 can exploit real vulnerabilities by reading security advisories," *The Register*, 17 April 2024, https://www.theregister.com/2024/04/17/gpt4_can_exploit_real_vulnerabilities/; Richard Fang et al., "LLM Agents can Autonomously Exploit One-day Vulnerabilities," *arXiv* (April 2024); Richard Fang et al., "LLM Agents can Autonomously Hack Websites," *arXiv* (February 2024).

that GenAI agents were able to exploit vulnerabilities by reading security advisories, thereby autonomously hacking websites. This research demonstrated the *potential* of agent-enhanced GenAI for autonomously writing malware, but it also highlights the challenges for evaluation. Without transparency about LLMs' training data, it is difficult to assess whether they can apply learned information to new, unseen scenarios and solve real-world problems. This uncertainty is further complicated when agents are equipped with search tools, blurring the lines between 'innate' capabilities and merely finding solutions online.

We have discussed how current GenAI systems perform against the goal of writing effective and operational malware. However, less sophisticated malware, which may not exploit zero-day vulnerabilities or employ the most up-to-date stealth techniques, could still influence the cybersecurity landscape due to the increased speed and scale of its development and deployment. Such malware (as can be found on the dark web) has been used to fine-tune systems such as WormGPT and DarkBERT.[45] These systems utilise models that have not been fine-tuned for safety, and are used as part of a subscription-based suite of cyberattack tools known as 'malware-as-a-service'.

Below we hypothesise some future scenarios based on the following (tentative) assumptions:

1) Limitations of training sets have been addressed.

2) Architectural improvements have been made to memorisation, attention and reasoning.

3) Models are now small enough they can be deployed as part of the malware, no longer requiring a centralised server.

[45] Youngjin Jin et al., "DarkBERT: A Language Model for the Dark Side of the Internet," *arXiv* (May 2023); Alp Cihangir Alsan, "Meet DarkBERT: Unravelling the Secrets of the Shadows," *Medium*, 10 August 2023, https://osintteam.blog/meet-darkbert-unraveling-the-secrets-of-the-shadows-26167e28a655.

Table 2. Advancements in system characteristics and future scenarios for malicious code generation

Characteristic/trait	Anticipated applications
Disguise	Techniques that allow the malware to alter its code when it executes (polymorphic[46]) or rewrite itself entirely (metamorphic) will improve the chances of the malware avoiding detection by signature-based security systems.
Tooling	Agents could write their own payloads, or create tools to overcome certain situations, e.g. a previously unseen file type, or privacy sub-system.
Planning	Agents could adapt their tactics in real-time, using the LLM for plan decomposition and task definition. This would allow the agents to work remotely with less need of direction.
Persistence	Multiple agents working in cooperation could provide a persistence technique, where if one is quarantined, the others would be able to learn and adapt from each other's experience.
Situational awareness	Agents could reason about the environment and adapt their communications to 'blend in', hiding in plain sight. This would increase the likelihood of circumventing behavioural detection systems.

Sophisticated malware often requires a delicate balance between stealth, security, and functionality. Such decisions require a level of tactical foresight and adaptive problem-

[46] Eran Shimony and Omer Tsarfati, "Chatting Our Way Into Creating a Polymorphic Malware," *CyberArk*, 17 January 2023, https://www.cyberark.com/resources/threat-research-blog/chatting-our-way-into-creating-a-polymorphic-malware; HYAS, *BLACKMAMBA: AI-synthesized, polymorphic keylogger with on-the-fly program modification* (HYAS: 2023), https://www.hyas.com/hubfs/Downloadable%20Content/HYAS-AI-Augmented-Cyber-Attack-WP-1.1.pdf.

solving that current LLMs do not possess,[47] and similar conclusions can be reached for GenAI's ability to autonomously find and exploit vulnerabilities.

To proactively monitor and detect the inflection points outlined in this section, the UK cybersecurity community should invest in an active research programme to monitor and track code generating systems' cyber-offensive knowledge, reasoning (contextual and interpretative understanding), and strategic decision-making (adaptability and continuous learning) capabilities. Cross-sector coordination is needed to ensure that large repositories of malware/exploits, malicious code, and vulnerability disclosures are securely managed.[48]

2.2 Radicalisation

The 'human-like' qualities of GenAI systems are the topic of considerable academic analysis. For example, a recent study found that people increasingly self-disclose to a robot over time, confiding more easily in them than humans.[49] Findings like this have clear implications for research on the persuasive and manipulative properties of GenAI systems, and by extension the radicalisation context.

This remains relatively under-researched in the AI safety landscape: the interim International Scientific Report on the Safety of Advanced AI described 'the use of conversational general-purpose AI systems to persuade users over multiple turns of dialogue' as an 'underexplored frontier'.[50] Current use cases of GenAI in the radicalisation context remain experimental and predominantly low-stakes, low-impact. Since a November 2023 report by Tech Against Terrorism (TAT), there does not appear to have been a significant evolution in the scale and nature of GenAI deployments for radicalisation. Predominant use cases include using AI art generators in messaging channels and for producing propaganda posters – TAT archived more than 5,000 pieces of AI-generated content produced by terrorist and violent extremist (TVE) actors.[51]

---

[47] Melanie Mitchell, "Can Large Language Models Reason?," AI Guide (Substack), 10 September 2023, https://aiguide.substack.com/p/can-large-language-models-reason; Jie Huang et al., "Large Language Models Cannot Self-Correct Reasoning Yet," *arXiv* (March 2024).

[48] Sarah Mercer and Tim Watson, "Generative AI in Cybersecurity: Assessing impact on current and future malicious software," *CETaS Briefing Papers* (June 2024).

[49] Guy Laban et al., "Building Long-Term Human-Robot Relationships: Examining Disclosure, Perception and Well-Being Across Time," *International Journal of Social Robotics* 16 (2024): 1-27; Zining Wang et al., "Ain't Misbehavin' – Using LLMs to Generate Expressive Robot Behavior in Conversations with the Tabletop Robot Haru," *arXiv* (February 2024).

[50] HM Government, *International Scientific Report on the Safety of Advanced AI: Interim Report* (Department for Science, Innovation and Technology: 2023), 43.

[51] 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.

Reasons for lack of adoption will partly depend on the specific *type* of TVE group being discussed. For example, numerous experts believe that TVE groups which operate in more traditional, hierarchical structures may be less willing to **cede control of a narrative** to unpredictable machines.[52] This is closely linked to the issue of **reliability** – the tendency for GenAI systems to hallucinate will compound concerns about message discipline. Some TVE groups put more of a premium on authoritative, single-source propaganda, whereas looser networks of extreme right-wing terrorist (ERWT) actors may not view the sanctity of the messaging in the same way and adopt a more pragmatic position regarding means and ends. In any case, the volume of real-world use cases is still too low to make a definite judgment on the variance across groups.

Moreover, there needs to be a commensurate **skills base** within a TVE group to consistently exploit GenAI capabilities. One interviewee suggested that terrorist groups will often prefer to upskill trusted individuals already known to the group rather than outsourcing to potentially untrustworthy individuals from outside, which could also explain slower adoption.[53]

One consistent factor across different types of TVE groups is their concern about **operational security** and the digital footprint associated with GenAI: GenAI systems are subject to monitoring which detect malicious usage and attempts to bypass in-built restrictions. One factor that could affect this is the pace of change in Edge AI and the trend towards running models locally/on-device.[54] The more confidence these groups have that their sensitive queries/prompts are untraceable by governments and law enforcement, the more experimentation there will likely be to understand possible use cases. In this context, the development of LLMs that can run on personal computers through 'quantization' is potentially significant.

As the radicalisation context often focuses on tailored individual interactions, there is an additional emphasis on **leveraging higher quality datasets**. Indeed, the more 'lawful' training data of GenAI systems creates a bias towards lawful answers, while there is also a bias towards western cultural and social norms in western-developed GenAI.

For a chatbot to be most effective in the 'radicaliser' role, it would first need information about the set of issues which the group cares about – here retrieval augmented generation (RAG) capabilities emerging alongside LLMs could play an interesting role. Layering an LLM on top of a large corpus of propaganda material is not technically difficult: 'if a terrorist

[52] Focus group with government representatives, 13 June 2024.
[53] Interview with government representative, 12 June 2024.
[54] Focus group with law enforcement representatives, 4 June 2024.

group with a huge library of things they have sent out in the past can make that more accessible for people to use as a knowledge base, it gives them more control over what the AI systems are putting out.'[55] Nonetheless, this does require internal data to be structured in a way that facilitates this extraction. Moreover, TVE groups would likely face challenges in hosting a vector store and retrieval pipeline which is geared to provide adequate improvements with RAG. It would require a considerable amount of storage and compute resource which would have to be achieved through third-party cloud providers implementing a certain level of monitoring.

A second aspect is needing more **specific information about the individual** being targeted. For example, if a GenAI system already possessed a high degree of understanding about the person it is interacting with at an early stage, its initial responses may be more engaging. Previous CETaS research has shown how 'language agents were able to portray believable personas (and) stay in character' with 'generated conversations rated highly for flow, pace, context and relevance, and other factors like the presence of filler words and consistent shared experiences and knowledge.'[56]

If this conversational capability can be matched with more tailored datasets, the appeal of GenAI systems for radicalisation would likely grow. A February 2024 study in Nature, reinforcing previous findings,[57] found that 'personalised messages crafted by ChatGPT exhibit significantly more influence than non-personalised messages' and that this was true across different domains of persuasion, including political ideology.[58] Another study showed that this effect is statistically significant even with small amounts of personal information being collected, indicating room for further improvement with the integration of fine-grained behavioural data.[59] However, there is also scepticism over whether micro-targeted messages are more persuasive than generic AI-produced content.[60]

---

[55] Interview with government representative, 12 June 2024.

[56] Ardi Janjeva, Alexander Harris, Sarah Mercer, Alexander Kasprzyk and Anna Gausen, "The Rapid Rise of Generative AI: Assessing risks to safety and security," *CETaS Research Reports* (December 2023): 81.

[57] Hui Bai et al., "Artificial Intelligence Can Persuade Humans on Political Issues," OSF preprints (February 2023); Alexis Palmer and Arthur Spirling, "Large Language Models Can Argue in Convincing Ways About Politics, But Humans Dislike AI Authors: implications for governance," *Political Science* 75, no.3 (April 2023); Josh Goldstein et al., "Can AI Write Persuasive Propaganda?" OSF preprints (February 2023).

[58] S.C. Matz et al., "The potential of generative AI for personalized persuasion at scale," *Scientific Reports* 14, no. 4692 (February 2024).

[59] Francesco Salvi et al., "On the Conversational Persuasiveness of Large Language Models: A Randomized Controlled Trial," *arXiv* (March 2024).

[60] Kobi Hackenburg and Helen Margetts, "Evaluating the persuasive influence of political microtargeting with large language Models," OSF preprints (August 2023); Almog Simchon, Matthew Edwards and Stephan Lewandowsky, "The persuasive effects of political microtargeting in the age of generative artificial intelligence," *PNAS Nexus* 3, no.2 (February 2024); https://arXiv.org/abs/2406.14508.

Table 3 below summarises some of the anticipated applications of GenAI systems within the radicalisation context. Section 3 will elaborate on the technological and systemic changes needed to see these become more commonplace.

Table 3. Advancements in system characteristics and anticipated applications for radicalisation

Characteristic/trait	Anticipated applications
Strategic advice	Assisting with fundraising and strategies for attracting donations; advising on routes to operational expansion, how to obfuscate malicious purchases and transferal of funds between donors and group members.
Targeted information gathering, scanning and vetting	Determining potential new recruits' suitability and authenticity. Synthesising up-to-date information on specific individuals for targeted approaches.
Creating and distributing content	Fully synthetic content generation, including speeches, images and interactive (gaming) environments, as part of one coherent chain.
Reinforcement and validation	Co-piloting or autonomously directing one-to-one, tailored radicalisation interactions which reinforce and validate lone-actor intentions.
Media spawning and variant recycling	Evading hashing algorithms through 'media spawning' and 'variant recycling'.

Analysing the role of technology on the process of radicalisation is a notoriously difficult challenge, as pathways to radicalisation are not linear, and there is variation in what triggers people to join a terrorist group and then a further set of factors which may trigger them to commit an act of violence. Moreover, human conversations tend to require 'nuanced handling of long-term contextual relationships and exhibit higher complexity through their

attention patterns', and there remains a significant gap in LLMs' ability to specialise in human conversations.[61]

But with the pace at which GenAI systems are advancing, there is ample reason to be concerned that this will be reflected in improved persuasive capabilities which provide an uplift to radicalisation efforts.[62] These may yield new manipulation tactics which humans are not prepared to combat.[63]

In this regard, one interviewee commented that 'with generative AI, what (we) thought would take a year to manifest is now happening far sooner', suggesting an awareness amongst practitioners that the pace of change is making it difficult to develop timely mitigations.

First, this indicates the importance of forecasting future inflection points which practitioners can develop mechanisms to track and manage today. Second, there is a need for clear feedback loops between the national security and law enforcement community which are responsible for detecting and combatting these applications downstream, and the AI evaluation community which is increasingly responsible for raising vigilance amongst developers.

2.3 Weapon instruction and attack planning

A common focus of the weapon instruction research landscape is the potential uplift in malicious actors' ability to plan and execute CBRN-related attacks.[64] The mechanisms for this may involve simplifying technical jargon, troubleshooting bottlenecks during weapons development, and increasing simulations for building bioweapons and executing attacks.[65]

These concerns have led stakeholders to commission safety evaluations for AI-assisted CBRN threats. However, results from these evaluations are mixed. A randomised control experiment conducted by RAND found no statistically significant difference between plans for bio-attacks developed with or without LLM assistance.[66] But researchers from Anthropic,

---

[61] Toshish Jawale et al., "Are Human Conversations Special? A Large Language Model Perspective," *arXiv* (March 2024).
[62] Daniel Siegel and Mary Bennett Doty, "Weapons of Mass Disruption: Artificial Intelligence and the Production of Extremist Propaganda," *Global Network on Extremism and Technology*, 17 February 2023, https://gnet-research.org/2023/02/17/weapons-of-mass-disruption-artificial-intelligence-and-the-production-of-extremist-propaganda/.
[63] Matthew Burtell and Thomas Woodside, "Artificial Influence: An Analysis Of AI-Driven Persuasion," *arXiv* (March 2023).
[64] Jonas B. Sandbrink, "Artificial Intelligence and Biological Misuse: Differentiating Risks of Language Models and Biological Design Tools," *arXiv* (December 2023); Daniil A. Boiko, Robert MacKnight and Gabe Gomes, "Emergent autonomous scientific research capabilities of large language models," *arXiv* (April 2023).
[65] Nathaniel Li et al., "The WMDP Benchmark: Measuring and Reducing Malicious Use with Unlearning," *arXiv* (May 2024).
[66] Christopher A. Mouton, Caleb Lucas and Ella Guest, *The Operational Risks of AI in Large-Scale Biological Attacks Results of a Red-Team Study* (RAND: January 2024), https://www.rand.org/pubs/research_reports/RRA2977-2.html.

partnering with leading biosecurity experts at Gryphon Scientific, found that leading GenAI models could produce 'sophisticated, accurate, useful, and detailed' knowledge, including harmful biological information at an 'expert level' for designing and acquiring biological weapons in 'some' cases.[67]

Similarly, Boiko et al.[68] found that AI systems demonstrated 'exceptional reasoning' in chemical experiments design, while Cem et al., conducting many-shot jailbreaking on Claude 2.0, found that the system could be prompted to elicit weapon instructions by combining multiple prompts.[69]

Most recently, UK AISI[70] found that five AI models (details not disclosed) had expert proficiency at generating knowledge about biology and chemistry. They found that on some topics such as advanced biology, some models outperformed experts by combining specific domain expertise with creative thinking, resulting in experimental approaches.

In some cases, the weapon instruction/attack planning domain is a natural extension of the malicious code generation and radicalisation use cases. For example, it is conceivable that a TVE group could use a GenAI system's coding capability to help with modifying video game graphics for violent artistic promotional material,[71] while deploying an LLM agent to gather the most relevant and up-to-date information about a recruitment target. Once initial contact is made, the group could run a fine-tuned chatbot 24/7 to interact with those previously identified individuals. If the group aims to carry out some form of real-world attack, it may then rely on the GenAI system to provide it with step-by-step experimental protocols and guidance for troubleshooting experiments.[72] The emergence of this **end-to-end 'AI attack chain'**, while unlikely in the short term, would undoubtedly transform the security landscape in the longer term.

There are three key elements to understanding the attack planning process.

---

[67] "Frontier Threats Red Teaming for AI Safety," Anthropic, last modified 26 July 2023, https://www.anthropic.com/news/frontier-threats-red-teaming-for-ai-safety.

[68] Daniil A. Boiko, Robert MacKnight and Gabe Gomes, "Emergent autonomous scientific research capabilities of large language models," *arXiv* (April 2023).

[69] Anil Cem et al., "Many-shot Jailbreaking," Anthropic research, April 2024, https://www-cdn.anthropic.com/af5633c94ed2beb282f6a53c595eb437e8e7b630/Many_Shot_Jailbreaking__2024_04_02_0936.pdf.

[70] Technical staff, "Advanced AI evaluations at AISI: May update," AI Safety Institute, 20 May 2024, https://www.aisi.gov.uk/work/advanced-ai-evaluations-may-update.

[71] Daniel Siegel and Mary Bennett Doty, "Weapons of Mass Disruption: Artificial Intelligence and the Production of Extremist Propaganda," *Global Network on Extremism and Technology*, 17 February 2023, https://gnet-research.org/2023/02/17/weapons-of-mass-disruption-artificial-intelligence-and-the-production-of-extremist-propaganda/.

[72] Jonas B. Sandbrink, "Artificial Intelligence and Biological Misuse: Differentiating Risks of Language Models and Biological Design Tools," *arXiv* (December 2023); Emily H. Soice et al., "Can Large Language Models Democratize Access to Dual-use Biotechnology?," *arXiv* (June 2023).

- The mindset (aligning with a set of views that would encourage acts of violence) and knowledge of how different types of attack are carried out.
- An intention to act on that knowledge.
- Possession of the materials or equipment needed to conduct the attack.

According to one interviewee, if at least two of these three factors are present, there will be cause for concern, and GenAI systems could play an important role in each of those stages.[73] Nonetheless, similarly to the previous analysis on radicalisation, there is limited current evidence of GenAI systems being used to assist in weapon instruction or attack planning. The reasons for this are also likely to be similar.

GenAI systems still struggle with **reasoning about the world** as humans experience it, and this is a significant shortcoming when it comes to attack planning. Tactical foresight, anticipating unintended consequences and mitigating trade-offs are all key components in this regard.[74] The same can be said for the **hallucinations** associated with GenAI systems – if the system provides incorrect technical or scientific information which a malicious actor relies upon, the risk of operational failure is very high. This consideration is especially prevalent with attack planning, where a group may only have one attempt to carry out a successful attack. Improvements in **multimodal GenAI systems** that can absorb information from images and videos for more holistic troubleshooting responses may alleviate some of these shortcomings.[75]

The fact that there is often a hard **skills and infrastructure barrier** may also heighten malicious actors' reluctance to trust the system's outputs:[76] buying into an ideology is fundamentally different from how one learns about chemistry or biology where there can be demonstrably incorrect answers. It remains unclear whether GenAI systems meaningfully uplift actors with synthesising, weaponizing and delivering a biological or chemical agent, relative to the Internet's capabilities: some studies suggest that equipping LLMs with specialised computational tools and the ability to search the web could shift the dial.[77]

Table 4 below summarises some of the anticipated applications of GenAI systems within the weapon instruction/attack planning context:

[73] Interview with government representative, 12 June 2024.
[74] HM Government, *International Scientific Report on the Safety of Advanced AI: Interim Report* (Department for Science, Innovation and Technology: 2024), 46.
[75] Julian N. Acosta et al., "Multimodal biomedical AI," *Nature Medicine* 28 (September 2022): 1773-1784.
[76] HM Government, *International Scientific Report on the Safety of Advanced AI: Interim Report* (Department for Science, Innovation and Technology: 2024), 46.
[77] Andres M. Bran et al., "Augmenting large language models with chemistry tools," in *37th Conference on Neural Information Processing Systems* (NeurIPS 2023: AI for Science Workshop).

Table 4. Advancements in system characteristics and anticipated applications for weapon instruction and attack planning

Characteristic/trait	Anticipated applications
Red teaming	Constant feedback and challenge on attack plans to improve lethality.
Weapon crafting	Using GenAI to help craft 3D-printed weapons.
Advanced OSINT gathering	Performing advanced OSINT gathering on specific targets.
Advanced summarisation and translation	Summarisation and translation of lengthy, technical or otherwise obscure documents.
Enabling VR/AR capabilities	Incorporating GenAI in illicit VR/AR environments.

One of the applications mentioned above – performing advanced OSINT gathering on targets – was also the subject of a previous CETaS case study.[78] This case study explored how an agent-based system (Llm_osint) could build a dossier on an individual and permit users to ask questions about them. The system first instructs the agent to conduct a web search curating an initial overview of the individual, then composes more tailored questions based on the information gathered and examines mentions of published work (if applicable). The agent can then be instructed to utilise this information for a nefarious task. This could provide a malicious actor with a detailed picture of a specific individual's day-to-day life, including social spaces they tend to frequent.

Use cases of this sort are an indicator of a potential future where malicious actors work in tandem with GenAI systems to ideate, develop, refine, perfect and then implement activities designed to cause significant public harm. The evaluation community must scan for and

[78] Ardi Janjeva, Alexander Harris, Sarah Mercer, Alexander Kasprzyk and Anna Gausen, "The Rapid Rise of Generative AI: Assessing risks to safety and security," *CETaS Research Reports* (December 2023).

monitor the inflection points that would make this future a reality, focusing on evaluating capabilities like prediction, adaptability and multi-modal processing. The national security and law enforcement community, which has the day-to-day responsibility for identifying and responding to the anticipated applications listed above as they emerge, must play a key role in ensuring those evaluation approaches are reflective of the evolving operational picture.

3. Primer for Evaluating Malicious AI Risk

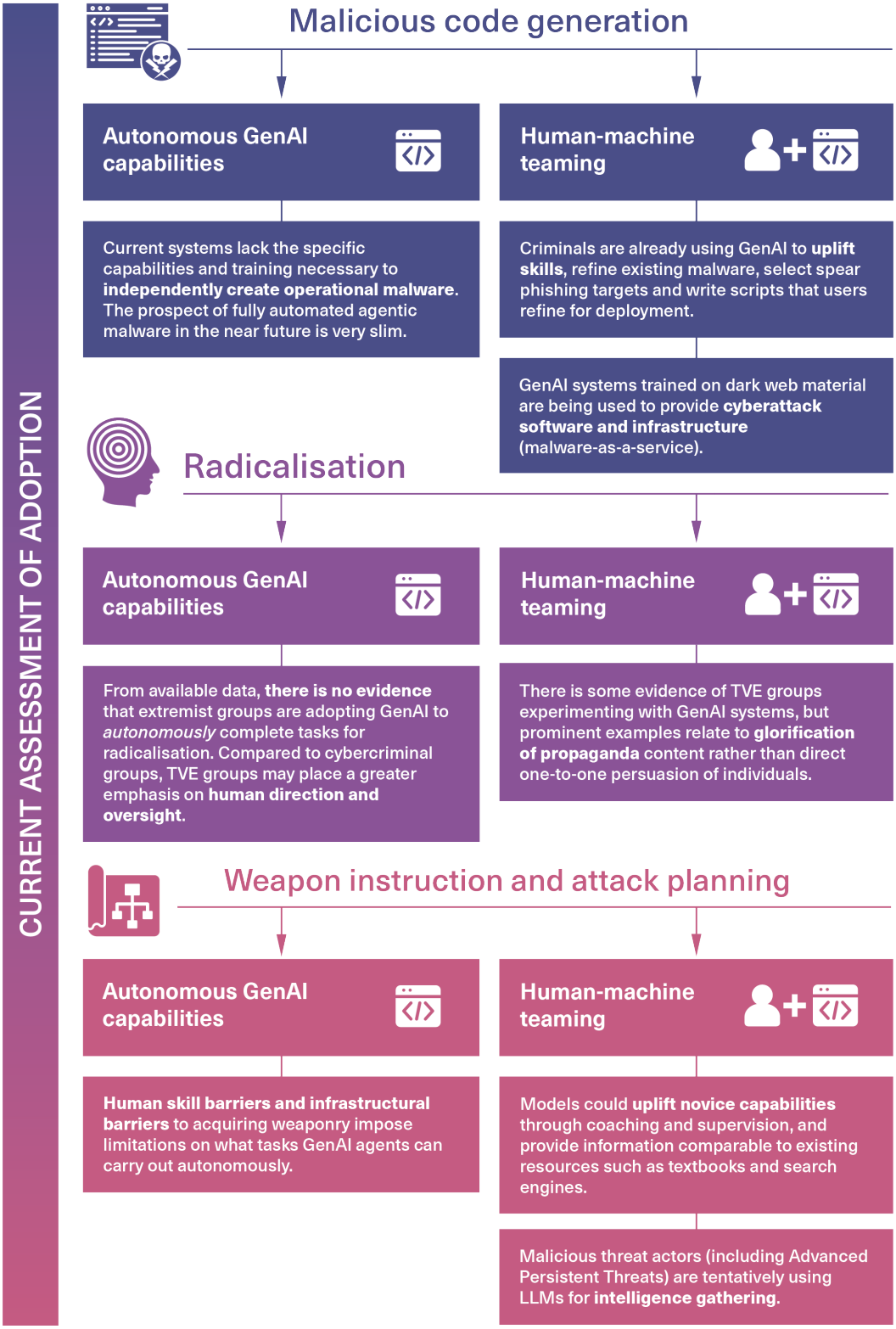
This final section explores the conditions that increase the likelihood and/or impact of risks occurring across the three threat areas of malicious code generation, radicalisation, and weapon instruction/attack planning. The aim is to use inflection points to structure the risk analysis process and help both those responsible for evaluating GenAI systems, and the national security and law enforcement community with understanding and prioritising potential risks.

As discussed in Section 1, a sociotechnical approach to evaluation is increasingly important. Human activity is adaptable, nuanced and context-driven, whereas computational algorithms often exhibit a rigid and brittle nature. Bridging the gap between technical solutions and the social requirements for GenAI deployment is central to designing an effective evaluation ecosystem.[79]

The figures and table below provide suggestions for how to evaluate the inflection points described in this paper, to inform proactive risk management strategies. For each threat domain, we present a current risk assessment, considering the autonomous GenAI context and the human-machine teaming context respectively. Then we outline a set of future inflection points of increased risk. Finally, we include a list of systemic/societal factors that would interact with – and potentially exacerbate – inflection points for that domain.

[79] Clark Barrett et al., "Identifying and Mitigating the Security Risks of Generative AI," *arXiv* (August 2023).

Figure 1. Current assessment of adoption[80]



[80] On malicious threat actors (including Advanced Persistent Threats) tentatively using LLMs for intelligence gathering, see: "Staying Ahead of Threat Actors in the Age of AI," Microsoft Threat Intelligence, Microsoft, last modified 14 February 2024, https://www.microsoft.com/en-us/security/blog/2024/02/14/staying-ahead-of-threat-actors-in-the-age-of-ai/.

Table 5. Future inflection points and mitigations/approaches to evaluation

	Future inflection points	Mitigations/approaches to evaluation
Malicious code generation 	1) If GenAI systems can **reason** and maintain a 'model' of code, they may execute better **tactical foresight** and **strategic decision-making** during cyberattacks.	*Reasoning:* a) Evaluating the model's ability to perform high-level cognitive tasks, using simulations and scenario-based testing. b) Examining the model's behaviour in a controlled environment to understand its decision-making patterns and limitations. *Code generation:* c) Evaluating complexity and novelty; measures of diversity of tactics, techniques and procedures employed by the outputs. d) Evaluating different models' strategies to avoid tripwires in different settings through red teaming.
	2) If GenAI enables autonomous agent architectures, agents could work in **cooperation** with each other, learning and adapting from each other's experiences.	a) Evaluating improved communication protocols for GenAI e.g. multi-agent dialogue systems.[81] b) Evaluating GenAI co-operative learning mechanisms e.g. multi-

[81] Debora C. Engelmann et al., "MAIDS – A Framework for the Development of Multi-Agent Intentional Dialogue Systems," in *AAMAS '23: Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems* (New York: Association for Computing Machinery, 2023), 1209-1217.

		agent learning algorithms or reinforcement with shared goals. c) Evaluating performance within virtual competitions (such as OpenAI's Gymnasium for reinforcement learning).[82]
	3) If highly capable GenAI systems become 'small' enough that they can be deployed as part of the malware, no longer requiring a **centralised server**, systems can run on smaller devices without trading off too much performance.	a) Cross-referencing existing accuracy evaluations with required resources, form factor and power usage; considering non-function metrics such as performance (speed) and reliability.
	4) If GenAI is trained on **high-quality examples** of sophisticated malware then offensive capability may significantly outperform defensive capabilities.	a) Monitoring the online availability of malware training datasets, including on the dark web. b) Issuing security guidance to repository owners of sophisticated malware/exploits and operationally viable vulnerabilities.
Radicalisation 	1) If the frequency of **GenAI hallucinations reduces**, and systems achieve **higher levels of social awareness and persuasiveness**, extremist groups who would typically be wary of delegating their messaging to unreliable	a) Measuring factuality of outputs using Wiki-FACTOR and News-FACTOR.[83] b) Measuring credibility of outputs using Frechet inception distance (FID) Scores or Inception Scores.

[82] Farama-Foundation, "Gymnasium," GitHub, July 2024, https://github.com/Farama-Foundation/Gymnasium.

[83] Dor Muhlgay et al., "Generating benchmarks for factuality evaluation of language models," *arXiv* (February 2024).

	machines could become more willing to do so. *Note: this inflection point is equally relevant for the attack planning context.*	c) Human feedback from domain experts to cross-validate responses. d) Using datasets like GlobalOpinionQA to test the ability of a model to reflect different opinions or cultural norms.[84] e) Behavioural experiments to understand if AI model outputs can shift beliefs.[85]
	2) Extremist groups may develop enhanced **RAG** capabilities on datasets tailored for radicalisation. There may be a proliferation of **custom knowledge bases** that GenAI systems can query to provide responses aligned with extremist messaging.	a) Determining and informing the monitoring standards of open-source RAG services and hosted vector stores (such as Llama Index). b) Understanding the technical and resource blockers facing TVEs attempting to acquire 'on-site' RAG capability.
	3) If GenAI could compile accurate information about **potential targets** for radicalisation, this would enable more tailored approaches and raise the likelihood of GenAI systems being used for scanning and vetting of new recruits/users.	a) Using output analysis to evaluate the amount of behavioural data (including consumer, demographic, psychological, geographic and engagement data) within a GenAI model.

[84] Esin Durmus et al., "Towards measuring the representation of subjective global opinions in language models," *arXiv* (April 2024).

[85] Hui Bai et al., "Artificial Intelligence Can Persuade Humans on Political Issues," OSF preprints (February 2023).

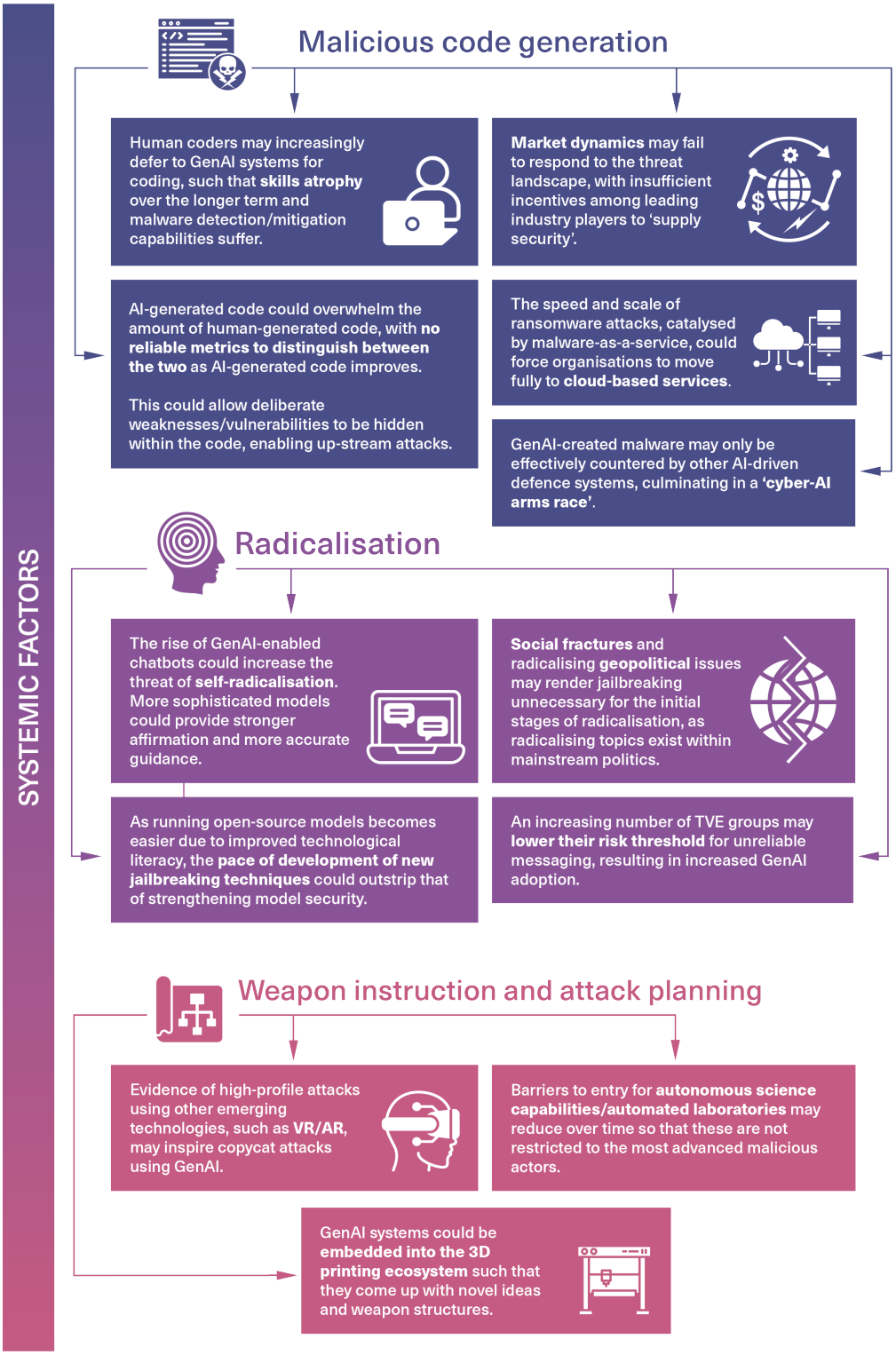
	4) Social media campaigns could be run by autonomous agents which **independently process exposed information**, utilising APIs and search engines, and integrating with social media accounts through open-source implementations such as AutoGPT and BabyAGI.[86]	a) Evaluating GenAI systems' segmentation accuracy (the ability to segment social media users) to enable better targeting of specific groups. b) Evaluating GenAI agents' ability to manage real-time data inputs and outputs, and manage and maintain social media activity and financial transactions.
Weapon instruction and attack planning 	1) If GenAI systems can adapt to the specific context of an attack plan, providing real-time tips and advice, this red teaming would improve attack lethality. **Multi-modal** features in future systems could go beyond text-to-text or text-to-image responses in an operational setting.	a) Evaluating GenAI's capacity for adaptability; determining how well the model can navigate new scenarios without human intervention. b) Evaluating GenAI's ability to handle multi-modal inputs using benchmarking approaches like 'MultiBench'[87] and the effect on generating consistent, correct and believable outputs.
	2) Malicious actors could move from using AI systems to help them gather information about the world to using them to make **predictions** which a prospective attacker acts upon.	a) Benchmarking of GenAI performance at temporal forecasting in the context of international events; compared to experienced human analysts.[88]
	3) If GenAI systems can be successfully integrated with	a) Evaluating techniques used to 'systematise' the outputs of

[86] Kai-Cheng Yang and Filippo Menczer, "Anatomy of an AI-powered malicious social botnet," *arXiv* (July 2023).
[87] Paul Pu Liang et al., "MULTIBENCH: Multiscale Benchmarks for Multimodal Representation Learning," in *NeurIPS '21: Proceedings of the 35th Conference on Neural Information Processing Systems* (New York: Association for Computing Machinery, 2021), https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/file/37693cfc748049e45d87b8c7d8b9aacd-Paper-round1.pdf.
[88] Chenchen Ye et al., "MIRAI: Evaluating LLM Agents for Event Forecasting," *arXiv* (July 2024).

	'narrow AI' tools, then natural language features of GenAI systems could make specialised tasks more accessible.[89]	GenAI such as structured generation, enabling easier integration with other sub-systems and services.
	4) If agent-based systems reliably identify connections with other linked individuals and can independently compile multiple dossiers, this would represent a leap in **OSINT gathering** on targets.	a) Adversarial testing and comparative analysis techniques to determine if a model's training data combined with agentic tools (i.e. search) yield sufficient information to enable more directed targeting.

[89] HM Government, *International Scientific Report on the Safety of Advanced AI: Interim Report* (Department for Science, Innovation and Technology: 2024), 47.

Figure 2. Systemic factors: potential future trajectories[90]



[90] Figure sources: interview with government representative, 7 June 2024; interview with government representative, 7 June 2024; focus group with law enforcement representatives, 4 June 2024; HM Government, *International Scientific Report on the Safety of Advanced AI: Interim Report* (Department for Science, Innovation and Technology: 2024), 47; focus group with law enforcement representatives, 4 June 2024.

In thinking about and forecasting malicious AI use, policymakers must appreciate that AI adoption will not solely depend on the properties of a given AI system, but also on an organisation's characteristics and readiness to adopt. This reality may be neglected outside of a sociotechnical approach: this demands more scrutiny on how technical, organisational and societal characteristics interact and mutually reinforce each other.

The figures and table above are an attempt to move beyond the technology-centric perspective on AI adoption that is dominant in the evaluation community, and encourage deeper understanding about the way that malicious groups work as part of the broader context of AI evaluation.

Conclusion

This paper aims to guide the evaluation of risk from GenAI across three threat areas: malicious code generation, radicalisation, and weapon instruction/attack planning.

It is important to highlight that evaluation as an endeavour has limits.[91] Firstly, evaluation approaches can only feasibly cover a subset of risks. Secondly, evaluation is value-laden, as the way in which the subset of risks is chosen and defined occurs through a series of normative decisions and expressions about what should be prioritised. Therefore, it is important that evaluation frameworks are seen as dynamic artefacts which evolve alongside technical and socio-cultural shifts.

Our work has also highlighted the importance of representing the expertise of the law enforcement and national security community in the way that the AI developer community thinks about risk and evaluations. No single discipline should define what risks are important, how they should be measured, and whether a system is safe. An interdisciplinary approach to GenAI evaluation in relation to national security is required to effectively capture the breadth of possible risks from technical system capability, human interaction with the AI system, and systemic factors.[92]

Common taxonomies and formalised channels are urgently needed to facilitate knowledge transfer between different stakeholders and disciplines, and this should be a high priority of the UK AI Safety Institute and Department for Science, Innovation and Technology (DSIT) in the months ahead.[93]

[91] Laura Weidinger et al., "Sociotechnical Safety Evaluation of Generative AI Systems," *arXiv* (October 2023).
[92] Seth Lazar and Alondra Nelson, "AI Safety on Whose Terms?" *Science* 381, no. 6654 (July 2023): 138.
[93] Laura Weidinger et al., "Holistic Safety and Responsibility Evaluations of Advanced AI Models," *arXiv* (April 2024).

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.

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.

Dr Sarah Mercer is a Principal Researcher working within the Defence and National Security Grand Challenge at The Alan Turing Institute. With 20+ years working within cybersecurity, 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 contributes to and provides engineering support to CETaS.

Tvesha Sippy is an Online Safety researcher within the Public Policy Programme at the Alan Turing Institute. Her background is in criminology and economics, and she uses survey methods and behavioural experiments to understand people's experiences of online harms, including deepfakes and other forms of AI-generated disinformation.

Centre for Emerging Technology and Security

BRIEFING PAPER