RAND EUROPE

THE CENTRE FOR
LONG-TERM RESILIENCE

Exploring red teaming to identify new and emerging risks from AI foundation models

Summary workshop report

Marie-Laure Hicks, Ella Guest, Jess Whittlestone, Jacob Ohrvik-Stott,
Sana Zakaria, Chryssa Politi, Cecilia Ang, Imogen Wade and Salil Gunashekar

For more information on this publication, visit www.rand.org/t/CFA3031-1

About RAND Europe

RAND Europe is a not-for-profit research organisation that helps improve policy and decision making through research and analysis. To learn more about RAND Europe, visit www.randeurope.org.

Research Integrity

RAND's publications do not necessarily reflect the opinions of its research clients and sponsors.

Published by the RAND Corporation, Santa Monica, Calif., and Cambridge, UK

Cover: Adobe Stock

Limited Print and Electronic Distribution Rights

Preface

On 12 September 2023, RAND Europe and the Centre for Long-Term Resilience organised a virtual workshop to inform UK government thinking on policy levers to identify risks from artificial intelligence foundation models in the lead up to the AI Safety Summit in November 2023. The workshop focused on the use of red teaming for risk identification, and any opportunities, challenges and trade-offs that may arise in using this method. The aims of the workshop were to:

- Explore red teaming as a method to identify new and emerging risks and capabilities of foundation models, how well it might work in practice, as well as its limitations
- Examine a spectrum of policy options to ensure red teaming can effectively identify new risks, including potential trade-offs and wider or unintended consequences
- Identify key considerations for policy development in assessing risks associated with foundation models, as well as further questions for research.

This report sets out a summary and overview of the discussions and findings from the workshop.

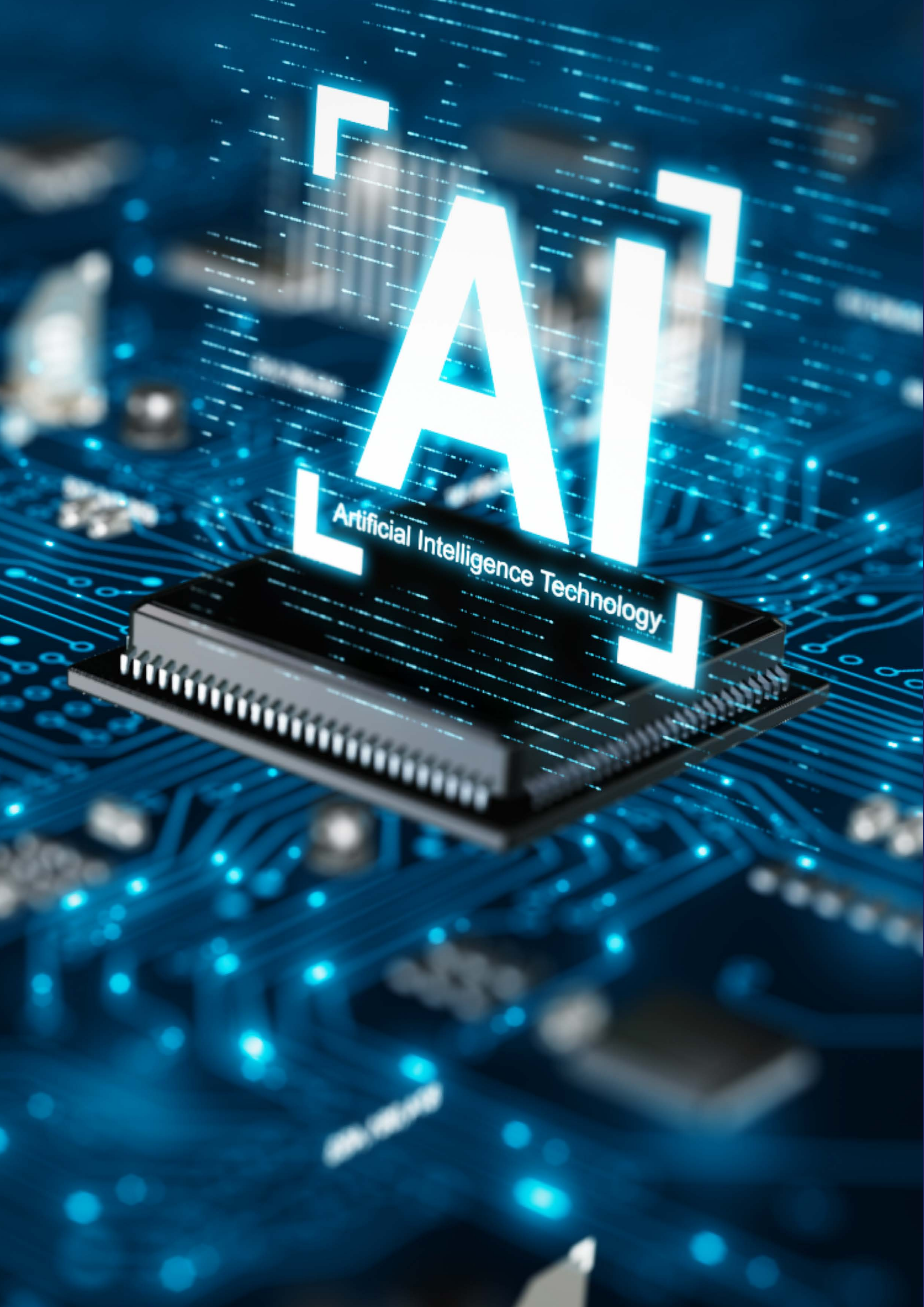
We would like to thank the workshop attendees for their participation and insights. We would also like to thank the quality assurance reviewers at RAND Europe, Susan Guthrie and Erik Silfversten for their critical review and feedback on the report. Finally, we would like to thank Jess Plumridge for helping to lay out the report and Ruby Russell for copy-editing. This work was funded by the RAND Corporation.

RAND Europe is a not-for-profit research organisation that aims to improve policy and decision making in the public interest, through research and analysis. Our clients include European governments, institutions, non-governmental organisations and firms with a need for rigorous, independent, multidisciplinary analysis.

The Centre for Long-Term Resilience is an independent think tank with a mission to transform global resilience to extreme risks. We do this by working with governments and other institutions to improve relevant governance, processes, and decision making.

For more information about RAND Europe or this document, please contact:

Salil Gunashekar (Deputy Director, Science and Emerging Technology)
RAND Europe
Eastbrook House, Shaftesbury Road
Cambridge CB2 8DR
United Kingdom
Email: sgunashe@randeurope.org



AI
Artificial Intelligence Technology

High-level summary

RAND Europe and the Centre for Long-Term Resilience hosted a virtual workshop on 12 September 2023 to discuss red teaming as a method to identify the risks and capabilities of artificial intelligence (AI) foundation models, how it might be implemented, as well as its limitations. The workshop sought to inform UK government thinking on different policy levers to identify risks from broadly capable AI models in the lead up to the AI Safety Summit in November 2023.

The workshop brought together a range of participants from across academia and public sector research organisations, non-governmental organisations and charities, the private sector, the legal profession and government. The workshop consisted of interactive discussions among the participants in plenary and in smaller breakout groups.

The views and ideas discussed at the workshop have been summarised in this short report to stimulate further debate and thinking as policy around this topical issue develops in the coming months.

The discussion focused on the following themes associated with the use of red teaming with AI foundation models to identify risks:

The term 'red teaming' is loosely used across the global AI community. A crucial first step is to develop a clear and shared taxonomy, along with shared norms and good practice around red teaming, for example, regarding who to involve, how to implement it and how to share findings.

Red teaming is one specific tool that is part of the wider risk identification, assessment and management toolbox. It is not a governance mechanism in itself.

Red teaming is useful in certain cases, in particular medium-term risks and assessment of known risks. Key limitations of red teaming included identifying unknown or chronic risks.

The socio-technical aspect of red teaming – who does it and in what context – must be actively considered. Embedding a diversity of perspectives, with deep understanding of the risks, the domain, and the actors or adversaries, is essential to improve a red team's effectiveness.

Specific methods such as red teaming should not be the focal point of mandated risk-management activities. If mandates are put in place, they should instead focus on holistic approaches and risk-management frameworks.

1. **Introduction**

Artificial intelligence (AI) holds huge promise, but it also presents several pressing challenges and risks. There is a great deal of uncertainty around how the technology – and its governance around the world – will develop in the coming months and years. On 12 September 2023, RAND Europe and the Centre for Long-Term Resilience organised a virtual workshop to inform UK government thinking on policy levers to identify risks from AI foundation models in the lead up to the AI Safety Summit in November 2023.[1] The workshop focused on the use of red teaming for risk identification, and any opportunities, challenges and trade-offs that may arise in using this method. Managing or mitigating risks, once identified, was not within the scope of the workshop. The workshop brought together a range of perspectives with 26 participants from across academia and public sector research organisations (10), non-governmental organisations and charities (2), think tanks (3), the private sector (2), legal profession (1) and government (8).

Given the dynamic nature of AI, predicting its associated risks is challenging. With heightened interest from stakeholders globally in exploring the impacts and implications of foundation and frontier models (see definitions in Box 1), a variety of approaches to identify new and emerging risks are being explored, including red teaming, described in more detail in Box 2. While multiple risk assessment methods can be used with these models, we chose to focus on red teaming because it was identified in the recently published voluntary commitments from leading AI companies secured by the White House, and is being used by companies developing these models, as well as at noteworthy events such as the hacker conference DEFCON 31.[2],[3] For example, OpenAI recently launched an open call for the OpenAI Red Teaming Network.[4] Participants in the workshop also suggested other methods for risk identification, which are presented in this paper with suggestions for their further exploration.

This paper sets out a summary of discussions from the workshop with findings and reflections from the different activities.

[1] Department for Science, Innovation and Technology (2023).

[2] White House (2023).

[3] Rivera Campos (2023).

[4] OpenAI Red Teaming Network homepage (2023).

Box 1: Key definitions

Key terms were shared with participants to provide a common language and enable discussion at the workshop. Definitions are drawn from the literature, noting that a single term may be used differently in different communities. These definitions formed part of the discussion at the workshop, with nuance or alternative interpretations detailed in the workshop summary presented in this report:

Foundation models (FM) are AI systems that use machine learning models trained on large and broad data sets. Capabilities include a range of general tasks, such as text synthesis or image generation. Applications can be built on top of these models.[5]

Generative artificial intelligence encompasses AI systems that create new and original content (text, image, video, audio) based on user inputs such as text prompts.

Large language models (LLM) are foundation models that have been trained with large volumes of text data (billions of words).[6]

Frontier model is a term used to describe foundation models with new and cutting-edge capabilities. There is, as yet, no universally agreed upon definition. For example, the term may indicate that a new model carries potential risk or harmful capabilities that are yet to be identified. In other contexts, it is applied to any model with cutting-edge capabilities, including those that are currently available. In some cases, compute power or measures of computer performance, such as floating-point operations per second (FLOPs), are used to define a model as a frontier model.[7,8,9]

AI value chain or lifecycle refers to the steps and activities that contribute to creating a valuable product or service for end users – in this case, for AI systems. A single organisation may perform one or more of the steps. McKinsey defines the generative AI value chain as: computer hardware chips, cloud platforms, foundation models, model hubs, applications and services.[10]

Red teaming consists of adversarial activities carried out by a group or individuals on a system with the support of the system's host organisation or owner. This can include attempts to breach defences, compromise systems or achieve another malicious act. Vulnerabilities and risks can be identified and communicated to the system host/owner so they can be resolved to ensure safety and security in case of a genuine attempt by malicious actors.

Risk is defined as a hazard or threat, usually an acute event or chronic trend by likelihood and impact on society, security, people and/or infrastructure. Risk often carries a degree of uncertainty in likelihood, impact or both. The UK National Risk Register categorises risks across the following themes: terrorism; cyber; state threats; geographic and diplomatic; accidents and systems failures; natural and environmental hazards; human, animal and plant health; societal; conflict and instability.[11]

[5] Jones (2023).

[6] Jones (2023).

[7] Frontier Model Forum homepage (2023).

[8] Anderjung &Schuett & Trager (2023).

[9] Jones (2023).

[10] McKinsey Digital (2023).

[11] HM Government (2023).

Box 2: Red teaming and its application to FMs

History

Red teaming was developed by the US military during the Cold War as a strategic role-playing exercise between Soviet 'red teams' and American 'blue teams'.[12] It is widely used in the cybersecurity sector to assess the robustness of modern security systems against a wide array of attacks. Organisations employ red teams to attack their security infrastructure to identify vulnerabilities so they can be fixed before they are exploited by malicious actors. In this scenario, the organisation's security acts as a 'blue team' attempting to defend the system from attacks.

Application to foundation models

We regarded red teaming as a broad range of methods to probe and test AI systems in an adversarial manner to identify risks and produce harmful outputs, for the purpose of determining AI model capabilities and informing changes to the system to prevent such outputs. Leading foundation model developers already use red teaming to test and improve their models.[13] In the case of foundation models such as large language models, red teams attempt to prompt the model to produce undesirable text outputs. Red teaming can be used to identify a range of risks, from disinformation to information on designing biological weapons. Red teams often include risk-specific subject matter experts and technical experts with experience of 'jailbreaking' or engineering LLM prompts.

The White House Voluntary Commitments on Safe, Secure, and Trustworthy AI include a commitment to internal and external red teaming of models.[14] Internal testing is performed by employees within the company developing the AI system. External testing is conducted by people outside the organisation, who would likely have specific subject-matter or technical expertise.[15]

Red teaming can take place at different stages in the lifecycle of an AI system. Model developers conduct red teaming during model development and pre-release to identify and mitigate risks before the systems can impact people and society. Red teaming post-release can identify risks that only become visible once the models are widely used, and can enable continuous monitoring of previously tested risks to ensure mitigations remain effective.

[12] Office of the Under Secretary of Defense for Acquisition, Technology, and Logistics (2003).

[13] See, for example: OpenAI Red Teaming Network homepage (2023), Fabian (2023), Anthropic (2023), Meta (2023).

[14] White House (2023).

[15] External red teams may require some form of collaboration with the AI developers, particularly if they require forms of model access not available to the public. Other external red teams may act fully independently of AI companies and assess publicly available versions of the system.

2. **Workshop approach**

This section provides an overview of the workshop approach, including aims, structure and different activities and discussions as a frame of reference for the rest of this paper.

The workshop took place under the Chatham House Rule and was structured as a series of facilitated discussions and activities in plenary and breakout groups that aimed to:

- Explore red teaming as a method to identify new and emerging risks and capabilities of foundation models, how well it might work in practice, as well as its limitations
- Examine a spectrum of policy options (e.g. voluntary commitments, mandated disclosure with liability) to ensure red teaming can effectively identify new risks, including potential trade-offs and wider or unintended consequences
- Identify key considerations for policy development in assessing risks associated with foundation models, as well as further questions for research.

Ahead of the workshop, participants were invited to submit thoughts in writing via email or on a Mural board on (i) opportunities and challenges in identifying the risks and capabilities of foundation models and (ii) methods for risk identification. This exercise was method- and risk-agnostic. A short review of submissions for (i) and options for participants to add comments and thoughts formed the basis for the workshop warm-up activity (Section 3A). Non-red teaming risk identification methods identified by participants in their submissions are presented in Section 3G.

Participants were introduced to red teaming and split into two groups, focused on societal risks and security risks, respectively. Risks were broadly defined to allow for wide-ranging discussion. Security risks – such as FM-assisted bioweapon development or cyberattacks – were defined as a threat or risk involving a malicious actor. Societal risks – such as bias, discrimination, misinformation or intellectual property (IP) theft – were defined as having an impact on society and citizens. Groups were presented with three activities addressing: the suitability of red teaming to identify risks; red teaming across the AI value chain; and future-proofing the effectiveness of red teaming (Sections 3B-F).

The second half of the workshop took place in plenary and focused on exploring the spectrum of policy options to implement red teaming and identify risk. Provocation statements were used to prompt discussion detailed in Section 4.

3. Assessing red teaming as an option to identify new and emerging risks from foundation models

A. Opportunities and challenges in identifying risks and capabilities of foundation models

At the start of the workshop, participants were invited to consider wider opportunities and challenges to effectively identifying new and emerging risks and capabilities from foundation models (see Figure 1). Opportunities ranged from considering political and industry support and international co-operation to building on existing bodies of work. Challenges focused on such themes as transparency, uncertainty, perverse incentives and technical challenges.

B. Defining red teaming

Participants were introduced to red teaming and split into two groups to discuss how it could be used to identify security and societal risks, respectively.

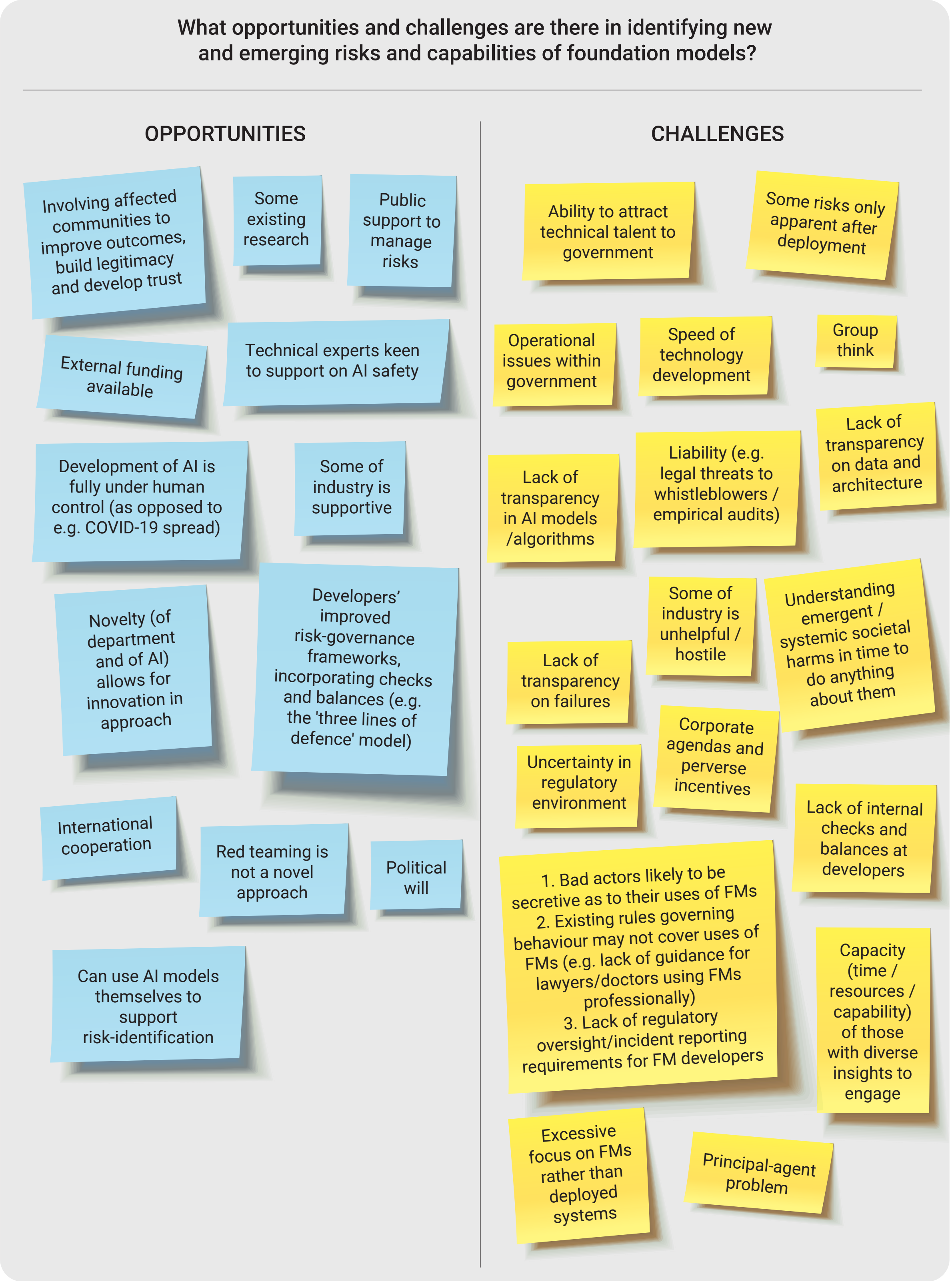
Discussion across both groups highlighted a need to clearly define what is meant by 'red teaming' in the context of foundation models and risk identification. Traditionally, red teaming is defined by the nature of the red team or adversary of concern and the system that needs to be protected or its value. When it comes to the application of red teaming to foundation models, participants noted that the term is sometimes used interchangeably with model evaluations, audits or impact assessments across a wide range of risks, in particular security and bias risks. However, participants agreed that there are distinct differences depending on the aims, timing and context of a red teaming activity. For example, participants did not necessarily consider the assessment of cyber security risks and computational access upstream in the AI value chain as red teaming, though it could be seen as such.

More generally, the term 'red teaming' has been used to broadly refer to any way of evaluating or testing a system for risk. In this context, we focused on a narrower definition of red teaming as a specific technique using a team that tries to prompt a system into revealing flaws or undesired behaviour.

C. Suitability of red teaming

Participants were asked to consider which risks could be identified through red teaming and which kinds of security and societal risks are likely to be missed by this method. The main themes and points of discussion from the groups are summarised in the Table 1 below.

Figure 1. Participant submissions for opportunities and challenges related to identifying new and emerging risks and capabilities from foundation models



What opportunities and challenges are there in identifying new and emerging risks and capabilities of foundation models?

OPPORTUNITIES

- Involving affected communities to improve outcomes, build legitimacy and develop trust
- Some existing research
- Public support to manage risks
- External funding available
- Technical experts keen to support on AI safety
- Development of AI is fully under human control (as opposed to e.g. COVID-19 spread)
- Some of industry is supportive
- Novelty (of department and of AI) allows for innovation in approach
- Developers' improved risk-governance frameworks, incorporating checks and balances (e.g. the 'three lines of defence' model)
- International cooperation
- Red teaming is not a novel approach
- Political will
- Can use AI models themselves to support risk-identification

CHALLENGES

- Ability to attract technical talent to government
- Some risks only apparent after deployment
- Operational issues within government
- Speed of technology development
- Group think
- Lack of transparency on data and architecture
- Lack of transparency in AI models /algorithms
- Liability (e.g. legal threats to whistleblowers / empirical audits)
- Some of industry is unhelpful / hostile
- Understanding emergent / systemic societal harms in time to do anything about them
- Lack of transparency on failures
- Corporate agendas and perverse incentives
- Uncertainty in regulatory environment
- Lack of internal checks and balances at developers
- 1. Bad actors likely to be secretive as to their uses of FMs
2. Existing rules governing behaviour may not cover uses of FMs (e.g. lack of guidance for lawyers/doctors using FMs professionally)
3. Lack of regulatory oversight/incident reporting requirements for FM developers
- Capacity (time / resources / capability) of those with diverse insights to engage
- Excessive focus on FMs rather than deployed systems
- Principal-agent problem

Table 1. Key points arising from two breakout-group discussions considering which risks may be identified through red teaming

Security risks	Societal risks
Applicability Participants perceived red teaming as effective for assessing certain types of risk, such as whether and to what extent foundation models could enable acquisition of knowledge and resources for bioweapon design and deployment. In this regard, red teaming was seen as helpful for identifying risks stemming fairly directly from the model itself – e.g. bias or lack of robustness – especially when assessing a specific potential risk. Effective red teaming must take a socio-technical perspective based on deep understanding of adversaries, be they hostile state or non-state actors. The composition of the red team is crucial as it needs to understand both the risk and the threat actor to explore what is realistically possible. Red teaming was viewed as effective to identify medium-term risks such as bioweapon development capabilities, but less effective for short-term or long-term risks. One participant noted that 'medium term' is contextualised here by the timescales for model training and capability development: for example, medium-term in this context might mean 2025.	**Applicability** Participants noted that red teaming can measure the propensity of a model to produce harmful outputs or contribute to harmful impacts but has limited ability to forecast or predict consequences for society at large. For example, with an application like ChatGPT, red teaming has been applied to explore how a user could jailbreak the model to bypass refusals and get the model to say what they want. Red teaming may enable probabilistic risk estimation for certain risks. For example, red teaming could estimate the likelihood or frequency of jailbreaking in certain contexts. Participants felt that red teaming might be better suited to more tightly defined risks (e.g. with known metrics and sources, such as discrimination).
Gaps Participants considered red teaming to be ineffective or of limited use in identifying short- and long-term security risks, risks of capabilities unknown to the developer (in the case of internal red teaming) and unknown unknowns.[16] Red teaming could be considered a tool for predicting what could happen in a particular scenario. As such, unknown unknowns are unlikely to be identified through red teaming. Similarly, the focus on adversarial attacks may limit red teaming's ability to identify risks not arising from misuse.	**Gaps** Participants highlighted that red teaming would be less effective for identifying risks stemming from how a model interacts with society. Red teaming faces a real challenge in this regard as many societal risks often cannot be anticipated, or may develop progressively through application and human use. The relationship between the model and user behaviour evolves interconnectedly over time. The dynamics of radicalisation, for example, are challenging to understand at different scales for anticipatory testing.

[16] 'Unknown unknowns' capture risks or potential consequences and impacts that are not predicted or foreseen.

Red teaming presents technical and feasibility characteristics that contribute to opportunities and limitations relevant to the identification of both security and societal risks, as summarised below.

Cross-cutting technical opportunities relevant to security and societal risks:

- A body of work and standards is available to build on: for example, bias and red teaming with ISO IEC DTS 12791[17]
- There are ways and approaches to model uncertainties in simple and complex systems that could be useful for red teaming: for example, Decision-Making Under Deep Uncertainty (DMDU) techniques can be used to grapple with uncertainty that cannot be quantified.[18]

Cross-cutting technical limitations relevant to security and societal risks:

- Red teaming has limited ability to identify unknown unknowns and risks that have not yet been predicted. Unpredictability across all risk types arises from the chaining of different AI and software systems, making unknown unknowns a particular challenge to address in risk identification
- Providing access to red teams, in particular external red teams, can be a barrier when the model is commercially sensitive
- Gaps in the technical understanding of foundation models remain, even among developers. Without fully understanding a model and its capabilities, red teaming is inherently limited in what it might uncover.

D. Red teaming across the AI value chain

Participants in the two breakout groups (on security and societal risks, respectively) were presented with a simple depiction of the AI value chain (computer hardware chips, cloud platforms, foundation models, model hubs, applications and services)[19] to consider at which stage of the value chain, for what risks, and how (e.g. internal or external) red teaming might be applied to identify risks from foundation models. It is worth noting that across both groups, participants raised comments on the value chain depiction itself with clarifications and reference to more fine-grained depictions. Further work to develop a shared understanding and representation of the AI value chain from a risk perspective would be valuable to both policymakers and the wider AI community. Table 2 sets out the key points from discussion for both risk categories.

[17] ISO (2023).

[18] RAND Corporation (2023).

[19] McKinsey Digital (2023).

Table 2. Summary of breakout discussions on red teaming across the AI value chain for security and societal risks

Security risks	Societal risks
• Participants viewed continuous red teaming across the AI value chain as necessary to identify threats. • At the development or training stage, red team scenarios could include quantified compute budgets that are useful for exploring potential thresholds for new capabilities. For example, this could be useful to inform potential export controls and access controls at the cloud-provider level. The dataset could also be red teamed to ensure there is no data poisoning and identify data to exclude from training (e.g. biological data relevant to bioweapons). At this stage in the value chain, internal red teaming would be appropriate, noting that external reporting or required standards may still be valuable (e.g. companies could do the dataset red teaming themselves but have a requirement to show publicly that they have done it to an adequate standard). • Moving downstream to market release and application, internal and external red teaming are both important. The farther along the value chain, the more accurately risks can be assessed. However, this may not always be true as finetuning the model at the application stage introduces unknown unknowns and may affect risk levels compared to the original release. • From a security perspective, assurance against threats infiltration is needed at all stages, and interaction with other models, systems and institutions must also be considered. • More generally, managing the balance between compliance and not lagging behind competitors will be a key challenge. • External red teaming – with subject-matter expertise that understands threat actors and risk scenarios – was considered more important than internal red teaming. Participants did not view post-deployment external red teaming once the model is accessible to some stakeholders or the public as sufficient – especially compared to pre-deployment external red teaming for security risks. As a bare minimum, it was felt that external experts should be given good access to the model. However, there is currently very little or no protection for external access and external red teams risk their access being removed.	• Participants viewed a focus on security risks rather than societal risks as potentially more appropriate earlier in the value chain (e.g. pre-deployment and in earlier development). This is partly due to the challenge of identifying and defining potential societal risks prior to deployment for red teaming to examine prior (in contrast to some security risks). • At the training stage, a key concern was raised around IP and data theft. Red teaming may provide a mechanism to identify this risk. This is heavily dependent on whether the model is a black box. • Currently, industry red teaming happens close to public release or shortly after. There is a gap with regards to earlier red teaming, for example during model training, with merit in extensive model testing. This would enable identification of risks that could potentially be resolved during training – for example, issues around bias. Companies may be more resistant if this delays release. • Context-based risks are inherently more visible at the application stage. For AI safety risks and societal risks that are highly context-dependent, red teaming may be more appropriate later down the value chain. • A key question to better understand is the extent to which policies and standards vary between model developers, those who use and apply the models and Application Programming Interface (API) users. How red teaming is applied by different stakeholders across the value chain may impact the effectiveness of risk identification and assessment. For example, varying assumptions, standards or approaches may create gaps or missed opportunities for effective risk identification. • AI could be used to assess risks across the value chain, identifying patterns and propensity to feed back into the system. This could potentially anticipate risks based on system use with extensive data on how people are actually using it. An important caveat here is that while a form of AI-led red teaming may be helpful to explore a defined risk or have a confirmatory purpose, this may not be applicable to all risks. Risks such as misinformation depend on how they are presented and pinpointing and investigating models' compliance at application will be challenging.

E. Practical considerations for red teaming

Across the range of risks discussed, several key points emerged as practical considerations to ensure the effective use of red teaming applied to foundation models:

- **Red teaming is better suited to risk assessment than novel risk identification:** Participants noted that red teaming is not a tool designed to identify the risk itself (particularly in the case of novel risks); rather it is a tool that can help verify whether a risk is present and assess it. For example, red teaming can focus on a particular risk, such as bio-risks from foundation models, and explore whether foundation models provide information to develop bioweapons, how actionable the information is and how easily obtained. It may also be helpful for monitoring the effectiveness of risk-mitigation measures such as refusals. This is reflected in the challenges mentioned earlier regarding red teaming as an ineffective tool for surfacing unknown unknowns.
- **The aim and the decisions that red teaming outputs will inform need to be clear:** This ensures both that red teaming is designed to deliver suitable outputs and generates a degree of transparency and accountability to act upon findings.
- **The red team must be credible:** Different risks and models require red teams with specific socio-technical expertise, including an understanding of the risk, technical understanding of the models and understanding of users or malicious actors. Without this deep understanding and emphasis on the diversity and expertise of the red team, the robustness and credibility of findings will be under question.
- **Inclusive red teaming:** Inclusivity and openness may be key considerations here, especially for societal risks. However, consideration should be given to ensure populations are not exploited, building on good practice and learning lessons. One participant identified the Māori Language Data Collective as an interesting model to consider in structuring inclusive red teaming.[20]
- **Risks from disclosures or leaks must be considered:** A responsible approach to sharing the findings of red teaming, including with whom, will be important to ensure disclosures contribute to desirable outcomes and mitigate any risks attached to disclosure itself. This is a particular concern for security or societal risks where malicious actors may seek to exploit vulnerabilities identified through red teaming. Similarly, red teamers must be trustworthy to ensure secure model sharing.
- **Automation and standardisation of disclosures could help support probabilistic risk assessment** by increasing the quality and availability of data to assess the likelihood of a particular risk.
- **Tracking technology evolution and business model evolution will be crucial** to understand the drivers of technological change, the organisations developing foundation models and how risks related to the technology may evolve. Red teaming can then be appropriately targeted.

[20] Birhane et al (2022).

- **Intelligent customers and awareness of limitations:** Red teaming risks becoming a tick-box exercise, or providing an inflated sense of assurance, if customers – whether government officials or senior executives – are unaware of red teaming's inherent limitations.

F. Future trends and how they might impact effectiveness of red teaming

Participants discussed how future trends and developments might affect the effectiveness of red teaming: in particular, technology development, technology accessibility and a catch-all 'other' category. Figure 3 collates the opportunities (purple), challenges (blue) and uncertainties (green) identified and explored in discussion. Two prompts provided a starting point for discussion: the rapid pace of technological development, and open source as a potential challenge to the effectiveness of red teaming.

The societal risks group focused on discussing the nuanced implications of open source, highlighting the opportunities of greater accessibility and diverse red teaming, as well as the challenge of mitigating the risks of making all information available, and code modifiable. Participants highlighted two examples: the Stable Diffusion leak is an example of code that, once openly available, was manipulated to produce sexually explicit material;[21] and the leak of Meta's LLaMA model prompted discussion about potential misuse and opportunities for safety improvement from open access.[22] A balance may be needed, with consideration given to which parts of a model are open source, as well as how to ensure that accessible and inclusive read teaming allows diverse communities to participate without being exploited.[23]

The security risks group highlighted uncertainties that might affect red teaming as social norms and the regulatory landscape evolve. Risks may not be viewed as risks until regulatory uncertainty is resolved, or societal norms evolve. For example, mental health risks were not considered when social media platforms were first launched, and regulation and legislation have taken time to develop. Regulation for children's online safety was first promised in 2018 and the UK Online Safety Bill was introduced to Parliament in 2021, passing the final stage in September 2023 and soon to become law.[24] This underscored the importance of taking a socio-technical view of red teaming, rather than turning it into a purely technical exercise.

Both groups noted an opportunity to create a community of red teamers, upskilling a wide range of people to be able to participate in red teaming.

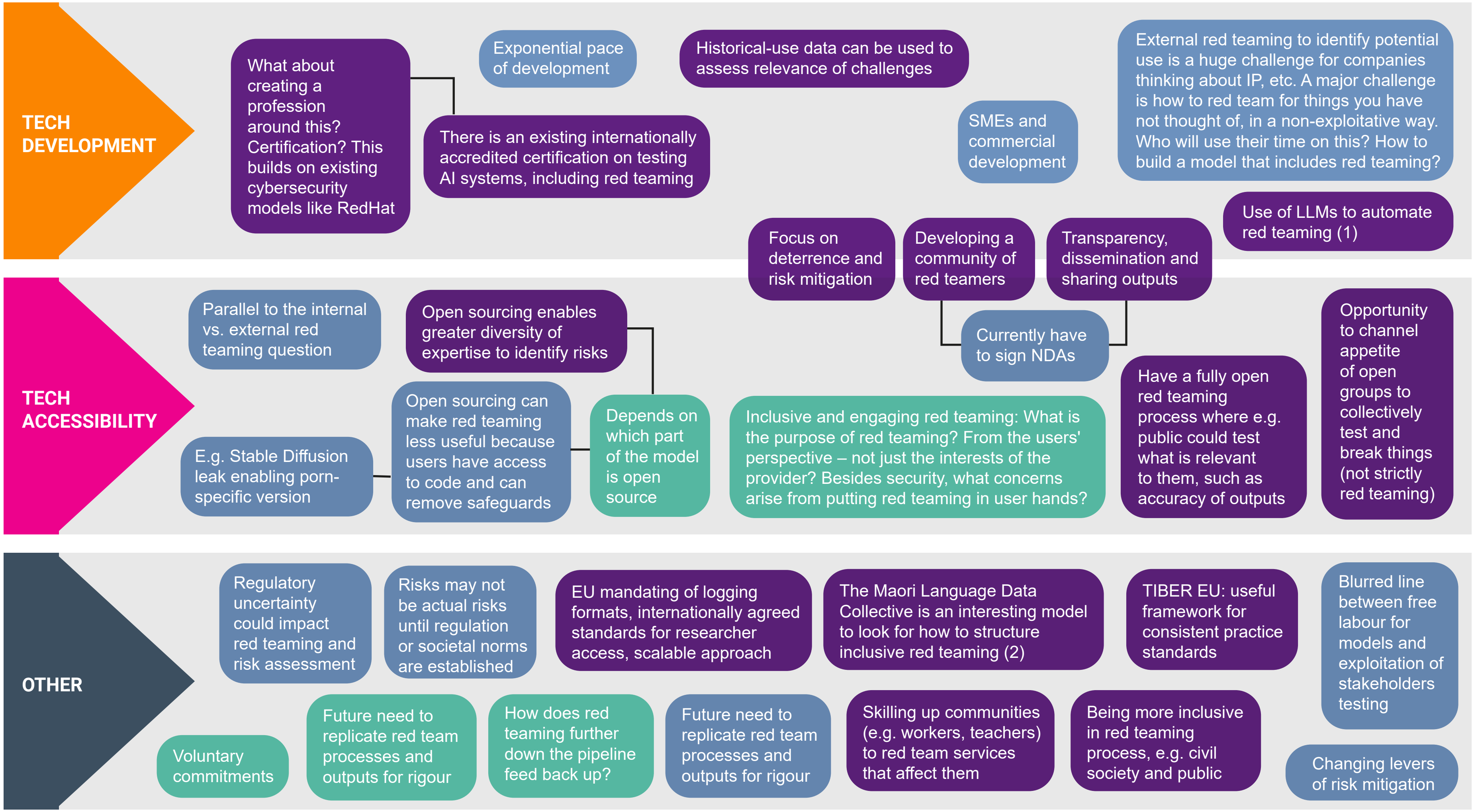
[21] Maiberg (2023).

[22] Vincent (2023).

[23] Seger et al (2023).

[24] NSPCC (2023).

Figure 3. Future considerations for red teaming (opportunities are depicted in purple, challenges in blue, and uncertainties in green)


TECH DEVELOPMENT
What about creating a profession around this? Certification? This builds on existing cybersecurity models like RedHat
Exponential pace of development
There is an existing internationally accredited certification on testing AI systems, including red teaming
Historical-use data can be used to assess relevance of challenges
SMEs and commercial development
External red teaming to identify potential use is a huge challenge for companies thinking about IP, etc. A major challenge is how to red team for things you have not thought of, in a non-exploitative way. Who will use their time on this? How to build a model that includes red teaming?
Use of LLMs to automate red teaming (1)
Focus on deterrence and risk mitigation
Developing a community of red teamers
Transparency, dissemination and sharing outputs
TECH ACCESSIBILITY
Parallel to the internal vs. external red teaming question
Open sourcing enables greater diversity of expertise to identify risks
Currently have to sign NDAs
Opportunity to channel appetite of open groups to collectively test and break things (not strictly red teaming)
E.g. Stable Diffusion leak enabling porn-specific version
Open sourcing can make red teaming less useful because users have access to code and can remove safeguards
Depends on which part of the model is open source
Inclusive and engaging red teaming: What is the purpose of red teaming? From the users' perspective – not just the interests of the provider? Besides security, what concerns arise from putting red teaming in user hands?
Have a fully open red teaming process where e.g. public could test what is relevant to them, such as accuracy of outputs
OTHER
Regulatory uncertainty could impact red teaming and risk assessment
Risks may not be actual risks until regulation or societal norms are established
EU mandating of logging formats, internationally agreed standards for researcher access, scalable approach
The Maori Language Data Collective is an interesting model to look for how to structure inclusive red teaming (2)
TIBER EU: useful framework for consistent practice standards
Blurred line between free labour for models and exploitation of stakeholders testing
Voluntary commitments
Future need to replicate red team processes and outputs for rigour
How does red teaming further down the pipeline feed back up?
Future need to replicate red team processes and outputs for rigour
Skilling up communities (e.g. workers, teachers) to red team services that affect them
Being more inclusive in red teaming process, e.g. civil society and public
Changing levers of risk mitigation

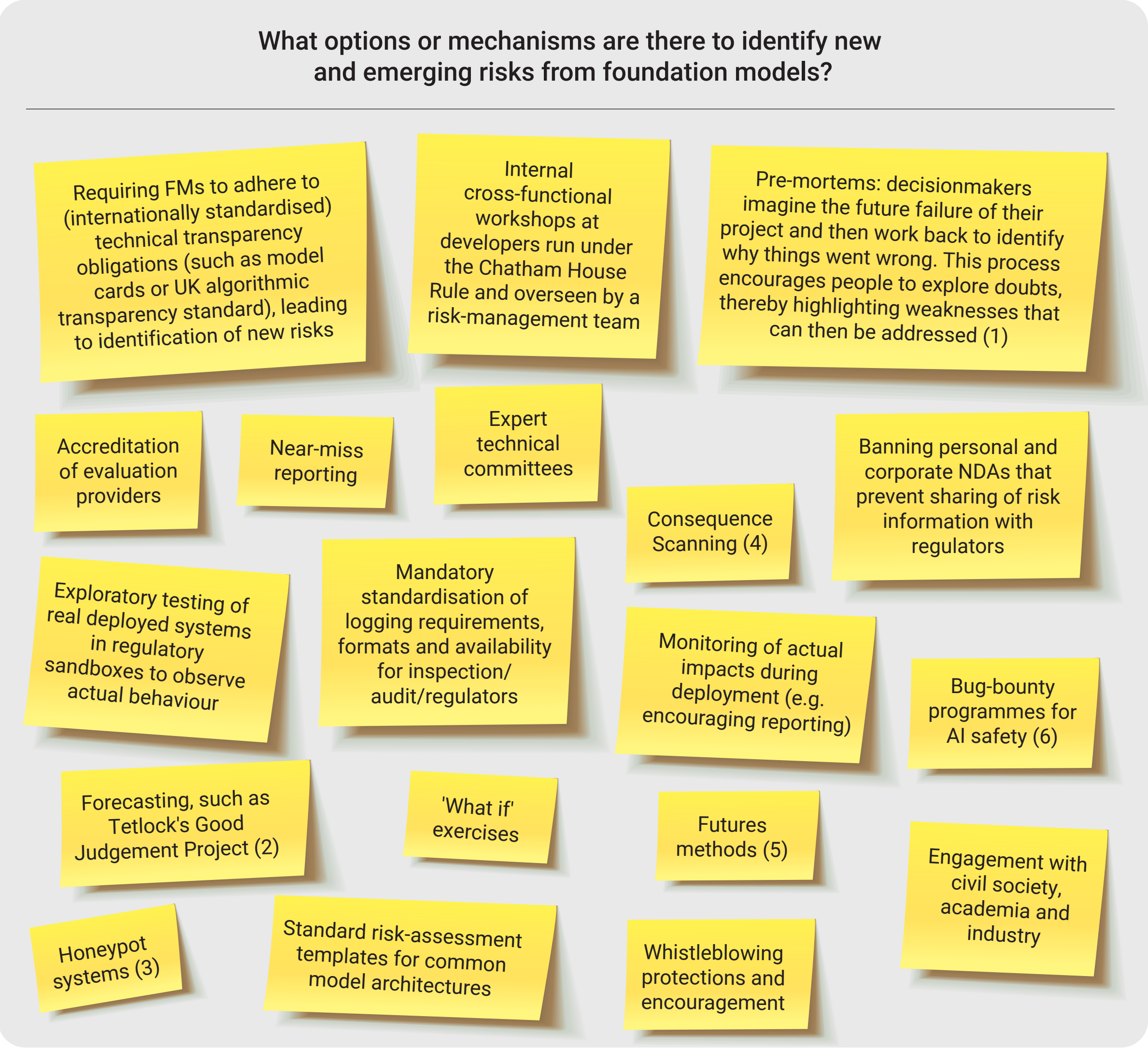
Note: Where participants provided additional references from the literature, these are captured with numbers in the post-its and listed in the footnote.[25]

[25] Participants in some cases provided references or links to further information including: (1) Perez et al (2022); (2) Birhane et al (2022).

G. Other risk identification methods

There was a consensus among workshop participants that red teaming alone is not sufficient to identify risks from foundation models. Participants identified other options and mechanisms to identify risks, illustrated in Figure 4. These were not discussed in depth during the workshop but point to the wider risk assessment toolbox that should be considered to assess risks from foundation models.

Figure 4. Options and mechanisms suggested by workshop participants to identify new and emerging risks from foundation models



What options or mechanisms are there to identify new and emerging risks from foundation models?

Requiring FMs to adhere to (internationally standardised) technical transparency obligations (such as model cards or UK algorithmic transparency standard), leading to identification of new risks

Internal cross-functional workshops at developers run under the Chatham House Rule and overseen by a risk-management team

Pre-mortems: decisionmakers imagine the future failure of their project and then work back to identify why things went wrong. This process encourages people to explore doubts, thereby highlighting weaknesses that can then be addressed (1)

Accreditation of evaluation providers

Near-miss reporting

Expert technical committees

Consequence Scanning (4)

Banning personal and corporate NDAs that prevent sharing of risk information with regulators

Exploratory testing of real deployed systems in regulatory sandboxes to observe actual behaviour

Mandatory standardisation of logging requirements, formats and availability for inspection/audit/regulators

Monitoring of actual impacts during deployment (e.g. encouraging reporting)

Bug-bounty programmes for AI safety (6)

Forecasting, such as Tetlock's Good Judgement Project (2)

'What if' exercises

Futures methods (5)

Engagement with civil society, academia and industry

Honeypot systems (3)

Standard risk-assessment templates for common model architectures

Whistleblowing protections and encouragement

Note: Where participants provided additional references from the literature, these are captured with numbers in the post-its and listed in the footnote.[26]

[26] Participants in some cases provided references or links to further information including: (1) Hallsworth et al. (2018); (2) Schoemaker & Tetlock (2016); (3) Wikipedia 2023; (4) doteveryone (2023); (5) Nesta (2023); (6) Wikipedia 2023.

4. Exploring policy options to incentivise red teaming and ensure it is used effectively to identify new risks from foundation models

To explore a spectrum of policy options – i.e. how red teaming could be supported or implemented – participants were presented with ‘what if?’ provocation statements. These statements were intentionally simplistic and crude to prompt discussion around opportunities and challenges that may entail a given policy option. This type of activity can help identify a policy option’s potential unintended consequences and practical considerations for its implementation to achieve a desired policy outcome. This workshop explored a spectrum of policy options through three provocations, starting with voluntary commitments, followed by standards and then mandatory disclosures with liability.

Alongside discussion of opportunities and challenges, a form of ‘Small to medium sized enterprise (SME) test’ was introduced with a prompt to encourage participants to consider SME-specific implications where relevant.[27] SMEs are key stakeholders to consider in policy design, especially as they may have limited resources to navigate and engage with policy development and may have specific requirements to manage challenges or opportunities.

A. What if red teaming was a voluntary commitment?

Opportunities with this policy option

- Participants highlighted that voluntary commitments would enable speed, flexibility and agility for implementation, without bureaucratic burdens compared to other policy options. An aspect of flexibility here would be that verification processes could be defined based on application need, rather than prescribed. Implementing red teaming for continuous testing of some risks, which is necessary given the dynamic nature of foundation models, would provide agility. One participant noted that because several general elections are coming up in the next one to two years (e.g. in the US, the UK, India), speed of action is important to address the risk of misinformation and disinformation.
- In a globally competitive landscape, some participants said a voluntary commitment would not disincentivise businesses from setting up in the UK.
- Voluntary commitments could create market incentives for certification.
- Red teaming would provide a means of demonstrating compliance with good practice or key measures.

[27] OECD (2021).

- Participants assumed that the burden on SMEs would be less than with a mandatory option.

Challenges with this policy option

- Participants noted a lack of incentives to carry out red teaming and take action based on its outputs. There is uncertainty over the scope and applicability of red teaming and how broadly it should be applied (e.g. each sector a product is used in, every country). Organisations will face the challenge of balancing the time and cost of red teaming against return on investment, shareholder incentives and other competing priorities. If a product is identified as risky, the question will be 'what happens next?' and whether it will be sold.
- Some participants said that transparency may be at risk – or the very least, would not be incentivised. Suspicion that individual companies or organisations might be letting themselves off the hook by 'marking their own homework' could undermine public trust. Reporting requirements for transparency and accountability could be covered by auditing.
- The use of external red teams may be limited. For example, model owners may only provide restrictive API access, limiting what external red teams are able to do, or they may remove access.
- Variability in standards, definitions and lack of uniformity could make it difficult to understand how effective a given red teaming exercise has been.
- Participants pointed out that in this scenario, an external central body would be unlikely to build up relevant expertise because activities would remain within individual organisations. This could create limitations, particularly around knowledge exchange and the sharing of good practice within the AI community and with the public sector and other key stakeholders.

Practical considerations for policy implementation

- Participants mentioned that voluntary red teaming would need to be heavily incentivised. One option would be to encourage the use of red teaming to demonstrate compliance with good practice, becoming a point of competition or distinction for commercial organisations.
- The discussion highlighted the importance of involving employees and employers. Employer buy-in is necessary, and there is an opportunity as currently most major companies want to say they are complying with voluntary commitments. The private-public AI ethics work carried out by Singapore with the Veritas Initiative[28] was mentioned as one example of good practice. Involving employees also provides internal accountability, following the example of lawyers who are responsible for a firm as well as their individual professional obligations.
- To be effective, some participants felt that red teaming would need to be part of a risk governance or risk management framework building on good practice. One participant took the view that no specific testing type should be mandatory, with the focus instead on mandating a whole-system approach or risk-governance framework.
- Bringing in diverse and external expertise requires partnerships, as developing or recruiting such expertise internally is

[28] MAS (2021).

unlikely to be possible. Marginalised groups will question what is being provided to them, or whether they are contributing towards the commercial success of an organisation that has limited returns to them. Improving the system does not necessarily improve the experience.

- One participant highlighted that for frontier models, the possibility of unexpected capabilities raises the importance of pre-training risk assessment and for other foundation models, pre-deployment red teaming may be more relevant to incentivise or mandate.
- Participants acknowledged that internal and external red teaming are needed. However, it is not a binary solution: if internal employees do not consider certain problems, this is likely to impact the efficacy of external red teaming.

B. What if standards include requirements for red teaming?

Opportunities with this policy option

- Participants discussed how standards provide an opportunity to improve the consistency of red teaming practices, such as the types of prompts or risks tested for. Standards may be best suited to component parts of red teaming activities, ensuring robust process and allowing for tailoring to specific risks and aims. For example, reporting standards are useful to ensure consistency and transparency.
- There is an opportunity to grow a large UK AI assurance market, building on existing strengths in the services sector, such as with lawyers and professional services organisations.
- Mandated risk assessments carried out by external parties and covering all forms of risk could help bridge the gap between the two risk camps (near-term risks and long-term or existential risks).
- There is an opportunity to see AI as a ‘normal’ scientific field and consider regulating across sectors such as biology, which includes experiments and human testing for pharmaceuticals. To share the regulatory framework more fairly, a distinction could be made between something leaked and misused, and a model released on the market.
- There is already a body of work and standards to build on or apply to red teaming.

Challenges with this policy option

- Standards are not enforced. The existence of a standard does not guarantee compliance.
- Some participants noted that standards are most helpful when the experience and knowledge is available to define ‘what good looks like’. Whether we have reached a level of shared and robust understanding to achieve this in the case of risk and foundation models is questionable. As the technology develops, such definitions must ensure that standards remain relevant and appropriate.
- Developing red teaming standards for foundation models is challenging. Terminology and views vary across countries and risks are context dependent. Unless they focus solely on the process, standards are a blunt instrument. In the case of security risks, shared standards would require a shared understanding of threat actors. Consensus and shared standards may be possible within certain communities (e.g. Partnership on AI or

Frontier Model Forum),[29,30] noting the lack of diverse perspectives in these fora. Achieving international standards was viewed as much more challenging.

- Standards development processes themselves are not always inclusive, which may lead to missed opportunities or unintended consequences that could otherwise have been identified.
- Standards are not necessarily always helpful for clarity or consistency, as they can be open to interpretation and used for convenience.

Practical considerations for policy implementation

- Participants said international standards bodies might have an important role to play, particularly concerning downstream applications of foundation models and sector-specific considerations.
- Participants noted that standards alone do not necessarily go further than voluntary commitments towards achieving a desired policy outcome, as compliance and adherence are not monitored. Used in a wider regulatory or risk management framework, a key consideration would be ensuring sufficient skilled expertise to deliver high-quality auditing or monitoring to review processes and outputs as relevant. This could form part of reporting mechanisms to a regulatory body, if in place.
- Standards should not only apply to use cases when models are brought to market. This would concentrate pressure on SMEs, which are frequently active at the application stage of the AI value chain. It would also fail to address the risk of exfiltration or misuse of models before release. Upstream developers should therefore share the burden of regulation and risk management.
- Red teaming, audits and impact assessments must have standards, which are essential to ensure standardisation of practice for external reporting and confirmation of steps taken to evaluate and mitigate risks. In this context, it is important that auditors are independent and subject to oversight.
- Participants highlighted the importance of mechanisms to share knowledge and good practice across key stakeholders and AI companies in this evolving space.

C. What if mandatory disclosures from red teaming were required, and had civil or criminal liability attached?

Opportunities with this policy option

- Several participants highlighted that mandatory action could enable effective red teaming. For example, mandatory availability of APIs, which are not necessarily public, would enable external red teaming.
- Liability can incentivise good behaviour. For example, food regulation placed liability for harm to customers on retailers, who in turn pushed requirements for assurance upstream to producers.
- Mandatory disclosures on impacts on fundamental rights could be a mechanism for risk identification and reporting.

[29] Partnership on AI homepage (2023).

[30] Frontier Model Forum homepage (2023).

Challenges with this policy option

- Some participants noted that power is concentrated upstream in the AI value chain. Caution should be taken not to place too much regulatory burden downstream, as this might be more challenging for SMEs to manage, and they are unlikely to be able to drive forward assurances from model developers.
- Enforcing requirements remains a challenge to achieving desired policy outcomes. Looking at the example of the 2008 financial crisis, participants noted that very few people were held criminally liable for a crisis with global impacts. Criminal liability is more frequently applied in cases of bodily harm or death.
- Disclosures carry risk in themselves. It will be important to ensure they are restricted to appropriate people and secure from malicious actors. A question was raised over whether failure to disclose would be criminalised, and what the implications might be.
- External reporting processes can create perverse incentives. Employees carrying out red teaming and risk assessments should not have to choose between telling senior leadership bad news or covering up.
- Concern was raised over SMEs engaging in high-risk activities. Regarding data protection, for example, the focus of concern may be further downstream (i.e. individual workers may pose greater risk to data security/privacy than an organisation's software). This could fall through the net if mandated disclosures or liability are focused upstream.
- Civil and criminal liability come into play post-harm.
- Regarding competition, the costs and complexity of development are already main entry barriers for foundation models. Participants therefore felt that regulatory burden would not be a key barrier to entry. This may be different at the application stage of the value chain, where they are more competitors.

Practical considerations for policy implementation

- Mandated action should focus on a wider ecosystem or outcomes, involving audits and different stakeholders, rather than specific tools or mechanisms to ensure risk is appropriately managed. This should include pre-harm measures, with accountability to take steps forward and have good processes in place for risk assessment and risk management. Regulatory sanctions or fines could form part of enforcement for pre-harm interventions and civil or criminal liability would capture harms that are not prevented ahead of time. For example, legislation could require that a system or model cannot be used to create bioweapons. This could include requirements for developers to demonstrate that the model cannot be used in such a way, and create imperatives for appropriate risk mitigation actions to be taken (rather than specifying an approach) through red teaming or other risk management mechanisms – concerning how developers market the model to downstream users, for example.
- Participants pointed out that liability should consider whether the organisation has carried out proper due diligence, such as red teaming activities, and provide exceptions accordingly. This would incentivise good practice and contribute to proportionate liability. Liability measures would likely be most effective if placed upstream, where power is concentrated.

- It will be important to understand where to place liability to incentivise good behaviour across the whole value chain. This could also provide a mechanism to manage the regulatory or liability burden, focusing on a smaller proportion of the value chain but ensuring good practice and desirable outcomes across its length. The right processes need to underpin both accountability and liability.
- Some participants said near misses should be captured. If policy comes into force post-harm, near misses and associated risky practice may not be captured, acted upon or learnt from. Consideration will need to be given to reporting near misses, including to whom, as part of risk management, legislative or regulatory requirements.
- Depending on the risk, liability could focus on harm or process. For example, if the risk is deemed as unacceptable, as with a system that could be used to develop bioweapons, liability would be best centred on ensuring processes such as red teaming are carried out. In cases where risk is deemed unacceptable, liability post-harm would not necessarily support the achievement of desirable outcomes, i.e. the risk not materialising.
- Proportionality of enforcement to risk is an important discussion. For example, voluntary commitments may be appropriate where the risk is lower, while stricter enforcement may be necessary for higher levels of risk. In practice, the difficulty of foreseeing which risks will be the most serious or harmful makes this challenging. What is non-catastrophic to one person can be catastrophic for another – one example being the spread of deepfakes.[31,32]

[31] Deeptrace (2019).

[32] Equity (2023).

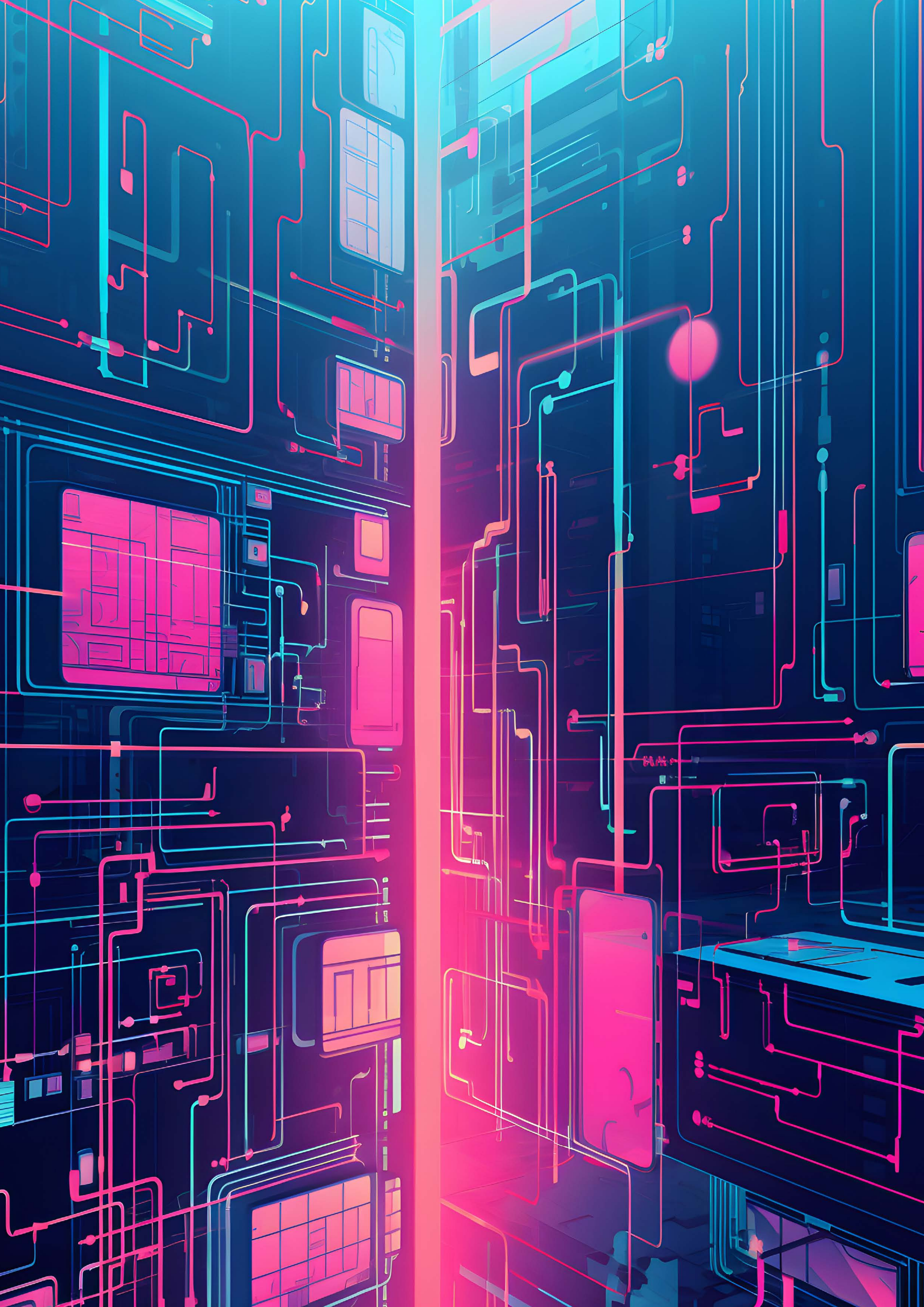
5. Reflections

Workshop participants viewed a whole-ecosystem approach to risk identification as best-suited to ensuring risks from foundation models – and AI more widely – are effectively identified and mitigated. It is useful to think of red teaming as one specific tool within the wider risk identification and risk assessment toolbox, rather than as a comprehensive governance mechanism in itself. Many participants did not view foundation or frontier models as the most appropriate system boundary, noting that narrow AI also has potential for high-risk applications in specific sectors such as biology.

Participants noted that specific methods such as red teaming should not be the focal point of mandated risk management activities. Rather, if mandates are put in place they should focus on holistic approaches and risk management frameworks. This includes bringing together diverse people with technical and risk-specific knowledge, resolving gaps in the technical understanding of AI systems (e.g. interpretability of models and measurability of risks) and understanding commercial and other incentives, as well as relationships and power dynamics across the AI value chain, to appropriately target policy measures that incentivise wider good practice.

In other words, the socio-technical aspect of red teaming – for example, who is doing it and in what context – must be actively considered. Embedding a diversity of perspectives, with a deep understanding of the risks, the domain, and the actors or adversaries, is likely to improve a red team's effectiveness. Governance approaches will need to acknowledge that models try to replicate human experience, making the range of harms broader and specific to context, region and language. Other key considerations include standards and good practice for the process itself, and use of independent, external red teams. To build buy-in and a community of practice, pragmatism regarding where the people with required skills and expertise are, and where they want to be, will be important.

The term 'red teaming' is used loosely across the AI community. A crucial first step is to develop a clear and shared taxonomy, along with shared norms and good practice, including who to involve in red teaming, how to implement it and how to share findings. However, it is worth noting that from an AI practitioner perspective, the trend of increasing red teaming represents a culture shift in itself. Whereas before there may not have been a deep concern about risks and impacts, this is changing and provides an opportunity to build on a growing appetite and support for AI safety and security.



References

Ajder, Henry, Giorgio Patrini, Francesco Cavalli & Laurence Cullen. 2019. *The State of Deepfakes: Landscape, Threats, and Impact.* Deeptrace. As of 18 October 2023: https://regmedia.co.uk/2019/10/08/deepfake_report.pdf

Anderjung, Markus, Jonas Schuett & Robert Trager. 10 July 2023. 'Frontier AI Regulation'. Centre for the Governance of AI. As of 18 October 2023: https://www.governance.ai/post/frontier-ai-regulation

Anthropic. 26 July 2023. 'Frontier Threats Red Teaming for AI Safety'. As of 18 October 2023: https://www.anthropic.com/index/frontier-threats-red-teaming-for-ai-safety

Birhane, Abeba, William Isaac, Vinodkumar Prabhakaran, Mark Diaz, Madeline Clar Elish, Iason Gabriel & Shakir Mohamed. 15 September 2022 19:20:13 UTC. 'Power to the People? Opportunities and Challenges for Participatory AI'. As of 18 October 2023: https://arxiv.org/abs/2209.07572

Department for Science, Innovation and Technology. 4 September 2023. 'UK Government Sets Out AI Safety Summit Ambitions'. As of 18 October 2023: https://www.gov.uk/government/news/uk-government-sets-out-ai-safety-summit-ambitions

doteveryone. 2023. 'Consequence Scanning – an Agile Practice for Responsible Innovations'. As of 18 October 2023: https://doteveryone.org.uk/project/consequence-scanning/

Equity. 2023. 'Stop AI Stealing the Show'. As of 18 October 2023: https://www.equity.org.uk/campaigns-policy/stop-ai-stealing-the-show

Fabian, Daniel. 19 July 2023. 'Google's AI Red Team: The Ethical Hackers Making AI Safer'. Google. As of 18 October 2023: https://blog.google/technology/safety-security/googles-ai-red-team-the-ethical-hackers-making-ai-safer/

Hallsworth, M., Egan, M., Rutter, J. & J. McCrae. 2018. Behavioural Government: Using behavioural science to improve how governments make decisions. The Behavioural Insights Team. As of 18 October 2023: https://www.bi.team/wp-content/uploads/2018/08/BIT-Behavioural-Government-Report-2018.pdf

Härlin, Tobias, Gardar Björnsson Rova, Alex Singla, Oleg Sokolov & Alex Sukharevsky. 26 April 2023. 'Exploring Opportunities in the Generative AI Value Chain'. McKinsey Digital. As of 18 October 2023: https://www.mckinsey.com/capabilities/quantumblack/our-insights/exploring-opportunities-in-the-generative-ai-value-chain

HM Government. 2023. 'National Risk Register 2023 Edition'. As of 18 October 2023: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1175834/2023_NATIONAL_RISK_REGISTER_NRR.pdf

International Standards Office. ISO/IEC DTS 12791 (under development). 'Information Technology – Artificial Intelligence – Treatment of Unwanted Bias in Classification and Regression Machine Learning Tasks'. As of 18 October 2023: https://www.iso.org/standard/84110.html

Jones, Elliot. 17 July 2023. 'Explainer: What Is a Foundation Model?'. Ada Lovelace Institute. As of 18 October 2023: https://www.adalovelaceinstitute.org/resource/foundation-models-explainer/

Maiberg, Emanuel. 22 August 2023. 'Inside the AI Porn Marketplace Where Everything and Everyone Is for Sale'. 404 Media. As of 18 October 2023: https://www.404media.co/inside-the-ai-porn-marketplace-where-everything-and-everyone-is-for-sale/

Meta. 18 July 2023. 'Meta and Microsoft Introduce the Next Generation of Llama'. As of 18 October 2023: https://about.fb.com/news/2023/07/llama-2/

Monetary Authority of Singapore. 3 March 2021. Veritas Initiative. As of 18 October 2023: https://www.mas.gov.sg/schemes-and-initiatives/veritas

Nesta. 2023. Futures. As of 18 October 2023: https://www.nesta.org.uk/feature/innovation-methods/futurescoping/

NSPCC. 19 September 2023. 'The Online Safety Bill Has Been Passed in a "Momentous Day for Children"'. As of 18 October 2023: https://www.nspcc.org.uk/about-us/news-opinion/2023/2023-09-19-the-online-safety-bill-has-been-passed-in-a-momentous-day-for-children/

OECD. 2021. 'The SME Test: Taking SMEs and Entrepreneurs into Account When Regulating'. As of 18 October 2023: https://www.oecd.org/gov/regulatory-policy/the-sme-test.pdf

Office of the Under Secretary of Defense for Acquisition, Technology, and Logistics. September 2003. 'Defense Science Board Task Force on the Role and Status of DoD Red Teaming Activities'. As of 18 October 2023: https://irp.fas.org/agency/dod/dsb/redteam.pdf

OpenAI Frontier Model Forum (homepage). 2023. As of 18 October 2023: https://openai.com/blog/frontier-model-forum

OpenAI Red Teaming Network (homepage). 2023. As of 18 October 2023: https://openai.com/blog/red-teaming-network

Partnership on AI (homepage). 2023. As of 18 October 2023: https://partnershiponai.org/

Perez, Ethan, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese & Geoffrey Irving. 7 February 2022 15:22:17 UTC. 'Red Teaming Language Models with Language Models'. As of 18 October 2023: https://arxiv.org/pdf/2202.03286.pdf

RAND Corporation. 2023. Robust Decision Making: Achieving Tomorrow's Goals Across an Uncertain Future. As of 18 October 2023: https://www.rand.org/pardee/methods-and-tools/robust-decision-making.html

Rivera Campos, Bianca. 17 August 2023. 'AI Safety at DEFCON 31: Red Teaming for Large Language Models (LLMs)'. As of 18 October 2023: https://www.giskard.ai/knowledge/ai-safety-defcon-31-red-teaming-llms

Schoemaker, Paul J. H. & Philip E Tetlock. May 2016. 'Superforecasting: How to Upgrade Your Company's Judgment'. Harvard Business Review. As of 18 October 2023: https://hbr.org/2016/05/superforecasting-how-to-upgrade-your-companys-judgment

Seger, Elizabeth, Noemi Dreksler, Richard Moulange, Emily Dardaman, Jonas Schuett, K. Wei, Christoph Winter, Mackenzie Arnold, Seán Ó hÉigeartaigh, Anton Korinek, Markus Anderljung, Ben Bucknall, Alan Chan, Eoghan Stafford, Leonie Koessler, Aviv Ovadya, Ben Garfinkel, Emma Bluemke, Michael Aird, Patrick Levermore, Julian Hazell & Abhishek Gupta. September 2023. 'Open-Sourcing Highly Capable Foundation Models – An Evaluation of Risks, Benefits, and Alternative Methods for Pursuing Open-Source Objectives'. Centre for the Governance of AI. As of 18 October 2023: https://cdn.governance.ai/Open-Sourcing_Highly_Capable_Foundation_Models_2023_GovAI.pdf

Vincent, James. 8 May 2023. 'Meta's Powerful AI Language Model Has Leaked Online – What Happens Now?'. The Verge. As of 18 October 2023:
https://www.theverge.com/2023/3/8/23629362/meta-ai-language-model-llama-leak-online-misuse

White House. July 2023. 'Ensuring Safe, Secure, and Trustworthy AI'. As of 18 October 2023: https://www.whitehouse.gov/wp-content/uploads/2023/07/Ensuring-Safe-Secure-and-Trustworthy-AI.pdf

Wikipedia. 'Bug Bounty Program'. 11 March 2014. As of 18 October 2023: https://en.wikipedia.org/wiki/Bug_bounty_program

Wikipedia. 'Honeypot (Computing)'. 4 August 2003. As of 18 October 2023: https://en.wikipedia.org/wiki/Honeypot_(computing)