Department for
Science, Innovation
& Technology

AI SAFETY
SUMMIT

HOSTED BY THE UK
1-2 NOVEMBER 2023

Emerging Processes for Frontier AI Safety

October 2023

Executive Summary

The UK recognises the enormous opportunities that AI can unlock across our economy and our society. However, without appropriate guardrails, such technologies can pose significant risks. The AI Safety Summit will focus on how best to manage the risks from frontier AI such as misuse, loss of control and societal harms. Frontier AI organisations play an important role in addressing these risks and promoting the safety of the development and deployment of frontier AI.

The UK has therefore encouraged frontier AI organisations to publish details on their frontier AI safety policies ahead of the AI Safety Summit hosted by the UK on 1-2 November 2023. This will provide transparency regarding how they are putting into practice voluntary AI safety commitments and enable the sharing of safety practices within the AI ecosystem. Transparency of AI systems can increase public trust, which can be a significant driver of AI adoption.

This document complements these publications by providing a potential list of frontier AI organisations' safety policies. These have been gathered after extensive research and will need updating regularly given the emerging nature of this technology. The safety processes are not listed in order of importance but are summarised in themes. The government is not suggesting or mandating any particular combination of policies – merely detailing the current suite available so that others can understand, interpret and compare frontier companies' safety policies. This document contains the world's first overview of emerging safety processes focused on frontier AI and is intended to be a useful tool to boost transparency. This conversation is for frontier AI and whilst it is important that safety is applied throughout the AI sector, it is also important that innovation is not stifled, hence why policies must be proportionate and based on capabilities which are the key driver of risk.

This document contains processes and associated practices that some frontier AI organisations are already implementing and others that are being considered within academia and broader civil society. It is intended as a guide for readers of frontier AI companies' AI safety policies to better understand what good policy might look like, though organisations themselves will be best placed to determine their applicability. Through this exercise, the government intends to help inform dialogue on potential appropriate measures for individual organisations to consider at the UK AI Safety Summit.

Pro-Innovation Approach

As demonstrated by our AI White Paper, innovation is at the heart of the UK's approach to this transformational technology. Only by gripping the risks at the frontier of AI development can we seize the opportunities for economic growth and public good. This exercise aims to consolidate various potential processes and associated practices that have been discussed and considered within the AI sector, civil society and by academics.

The important role that frontier AI organisations play in promoting the safe development and deployment of frontier AI will also benefit the wider AI ecosystem, including organisations that are not at the frontier. We are therefore keen to ensure that small to medium sized businesses are included in the conversation on AI safety as processes and practices continue to emerge.

Who is this document for?

This document has been written for leading AI organisations, those at the frontier of capabilities, as well as those who want to better understand their safety policies. It is therefore intended as a potential menu for a very small number of AI organisations at the cutting edge of AI development. While there may be some processes and practices relevant for different kinds of AI organisations, others - such as responsible capability scaling - are specifically developed for frontier AI and are not designed for lower capability or non-frontier AI systems. We invite frontier AI organisations to consider them in ways most appropriate to the risks posed by their specific models and the context in which they are developed and deployed.

This document has taken a targeted approach by focusing on frontier AI organisations; there are many AI organisations operating at a lower level of risk that are not expected to consider such a range of measures. This is in line with our proportionate and pro-innovation approach to addressing the risks of AI.

What is this document?

This document provides a snapshot of promising ideas and emerging processes and associated practices in AI safety today. It should not be read as government policy that must be enacted, but is intended as a point of reference to inform the development of frontier AI organisations' safety policies and as a companion for readers of these policies. The document consolidates leading thinking in AI safety from research institutes and academia, companies, and civil society, who we have collaborated and engaged with throughout its development.

Frontier AI safety is an ongoing project and processes and practices will continue to evolve through research and dialogue between governments and the broader AI ecosystem. There are valuable measures not yet considered, and some included in this document may ultimately prove technically infeasible or undesirable. This document is therefore not the final word on safety within frontier AI. We look forward to the AI ecosystem further developing these processes and practices as AI safety research evolves.

This document builds on progress that has already been made both within the UK and internationally to respond to rapid progress in AI development. In March 2023, the UK government published the AI Regulation White Paper, which set out our pro-innovation approach to addressing the risks associated with AI while ensuring its benefits can be realised. The request for companies to publish their AI Safety Policies and the publication of this supporting document demonstrates this flexible approach by focusing on frontier AI organisations with the highest risks, while recognising that - with technology progressing very quickly - processes and practice are still evolving.

Work on risks is also happening in other countries, international fora such as the OECD, GPAI, G7 and others, as well as through industry-led processes in standards development organisations. This document is intended to build on and complement these approaches and frameworks by setting out promising ideas and emerging processes and practices for frontier AI safety.

What is contained in the document?

This document sets out nine emerging processes and associated practices for frontier AI organisations' AI safety policies:

1. **Responsible Capability Scaling** provides a framework for managing risk as organisations scale capability of frontier AI systems. It enables companies to prepare for potential future, more dangerous AI risks before they occur, as well as manage the risks associated with current systems. Practices involve conducting thorough risk assessments, pre-specifying risk thresholds and committing to specific mitigations at each of those thresholds and being prepared to pause development or deployment if those mitigations are not in place.

Effective risk management in frontier AI will require processes across a range of risk identification and mitigation measures, including six that many leading companies committed to in July 2023:

2. **Model Evaluations and Red Teaming** can help assess the risks AI models pose and inform better decisions about training, securing, and deploying them. As new capabilities and risks can appear as frontier AI models are developed and deployed, evaluating models for several sources of risk and potential harmful impacts throughout the AI lifecycle is vital. External evaluation by independent, third party evaluators can also help to verify claims around the safety of frontier AI systems.

3. **Model Reporting and Information Sharing** increases government visibility into frontier AI development and deployment. Information sharing also enables users to make well-informed choices about whether and how to use AI systems. Practices involve sharing different information about their internal processes, safety and security incidents, and specific AI systems with different parties, including governments, other frontier AI organisations, independent third parties, and the public, as appropriate.

4. **Security Controls Including Securing Model Weights** are key underpinnings for the safety of an AI system. If they are not developed and deployed securely, AI models risk being stolen or leaking secret or sensitive data, potentially before important safeguards have been applied. It is important to consider the cyber security of AI systems, as well as models in isolation, and to implement cyber security processes across their AI systems, including their underlying infrastructure and supply chains.

5. **Reporting Structure for Vulnerabilities** enables outsiders to identify safety and security issues in an AI system. This is analogous to how organisations often set up 'bug bounty programs' for vulnerabilities in software and IT infrastructure. Practices include setting up a vulnerability management process that would cover many vulnerabilities - such as jailbreaking and prompt injection attacks - and have clear, user-friendly processes for receiving vulnerability reports.

6. **Identifiers of AI-generated Material** provide additional information about whether content has been AI generated or modified. This can help prevent the creation and distribution of deceptive AI-generated content. Investing in developing techniques to identify AI-generated content is important in a highly nascent field, as well as exploring the use of approaches such as watermarks and AI output databases.

7. **Prioritising Research on Risks Posed by AI** will help identify and address the emerging risks posed by frontier AI. Frontier AI organisations have a particular responsibility and capability to conduct AI safety research, to share their findings widely, and to invest in developing tools to address these risks. Collaboration with external researchers, independent research organisations, and third-party data owners will also be key to assessing the potential downstream social impacts of their systems.

Risk management will likely require further measures than those already committed to, however. We suggest two additional processes and associated practices:

8. **Preventing and Monitoring Model Misuse** is important as, once deployed, AI systems can be intentionally misused for harmful outcomes. Practices include establishing processes to identify and monitor model misuse, as well as implementing a range of preventative measures, and continually reviewing their effectiveness and desirability over time. Given the serious risks that misuse of frontier AI may pose, preparations could also be made to respond to potential worst-case misuse scenarios.

9. **Data Input Controls and Audits** can help identify and remove training data likely to increase the dangerous capabilities their frontier AI systems possess, and the risks they pose. Implementation of responsible data collection and cleaning practices can help to improve the quality of training data before it is collected. Careful audits of training data- both by frontier AI organisations themselves and external actors can also enable the identification of potentially harmful or undesirable data in training datasets. This can inform subsequent mitigatory actions, such as the removal of this data.

This document is for information only and should not be construed as providing any legal or other advice. Readers are encouraged to seek their own legal advice before implementing any AI safety policies.

Contents

Key terms

Below, we define a list of key terms referenced throughout the document. Specific technical terms are described within their relevant section.

AI or AI system: products and services that are ‘adaptable’ and ‘autonomous’ in the sense outlined in our definition in section 3.2.1 of the AI White Paper.

Dangerous capabilities: the abilities of an AI system to cause significant harm due to intentional misuse or accident.

Frontier AI: highly capable general-purpose AI models that can perform a wide variety of tasks and match or exceed the capabilities present in today’s most advanced models.

Frontier AI organisation: organisations developing frontier AI.

Relevant government authority: a national government body or intergovernmental organisation with the appropriate mandate to receive information from frontier AI organisations.

Safety: the prevention and mitigation of harms from AI.

Users: any person or organisation who uses a frontier AI system.

Responsible Capability Scaling

Summary

Decisions about whether and how to develop and deploy new AI systems are consequential. Deploying an excessively high-risk system may lead to significant harm from misuse or safety failures. Even just developing a high-risk system may lead to harm if it eventually leaks, is stolen or causes harm during internal deployment.

Responsible Capability Scaling is an emerging framework to manage risks associated with frontier AI and guide decision-making about AI development and deployment. It involves implementing processes to identify, monitor, and mitigate frontier AI risks, which include other processes and practices set out in this document and are underpinned by robust internal accountability and external verification processes.

We outline seven categories of practice regarding Responsible Capability Scaling:

1. **Conducting thorough risk assessments before developing or deploying new models, complemented with continual monitoring** (incorporating e.g. evaluation results, evidence from earlier models, and expert forecasts)

2. **Pre-specifying "risk thresholds" that limit the level of acceptable risk** (a level of risk of, for instance, cyberattacks or fraud that it would be unacceptable for a model release to create)

3. **Pre-committing to specific additional mitigations for systems at each risk threshold, followed by a residual risk assessment**

4. **Adapting mitigations to the relevant stages of deployment, recognising that models may be used in different ways and contexts than intended**

5. **Making preparations to pause development and/or deployment if risk thresholds are reached without the pre-agreed mitigations**

6. **Share details of risk assessment process and risk mitigation measures with relevant government authorities and other AI companies**

7. **Committing to robust internal accountability and governance mechanisms, alongside external verification** (e.g. internal risk governance, record-keeping, independent audits)

Background

As capability scales, many questions surrounding model development and deployment will warrant significant care. These questions include:

- what models to develop and how
- what level of security these models warrant during development
- whether and how to deploy a model, for instance whether it should be deployed through an API or open-sourced
- what datasets to use in training
- what guidance, if any, to provide to users
- what safeguards to be put in place.

Practices

1. Conduct thorough risk assessments before developing and deploying new models and as appropriate throughout the life cycle, complemented with continual monitoring

Risk assessments are valuable, as they can lead to responsible decisions - including decisions made deliberately in advance of capability advancements - and ultimately mitigate risk.

Develop rigorous risk assessment processes for models, which:

- Attempt to encompass all plausible and consequential risks from AI systems, including low-probability risks of severe harm
- Are informed by factors including, but not limited to:
  - Model evaluations and red teaming (see Model Evaluations and Red Teaming)
  - Evidence of previous models' impacts and capabilities
  - Knowledge of the latest research and developments in the field (see Priority Research on Risks Posed by AI)
  - Domain expertise both internally and externally
  - The results of data input audits (see Data Input Controls and Audits)
- Take into account the difficulty of producing reliable risk assessments, creating a culture that takes seriously the significant uncertainty underlying predictions of the risks associated with frontier models
- Take into account the potential benefits of the model
- Include learnings from information exchanges with industry, academia and government on the capabilities of comparable models

Devote resources to pre-development risk assessments, alongside prioritising pre-deployment risk assessments. Pre-development risk assessments are also important, because training a high-risk system can still lead to harm if the model is leaked, stolen, or otherwise unintentionally distributed.

Monitor systems both during development and after deployment. New risk assessments could be carried out in cases of fine-tuning or other substantial changes that could increase the danger of a model, such as the model gaining access to tools or plugins. This could occur alongside attempts to detect unexpected developments and new information that might have changed the results of the existing risk assessment, and which could also trigger a new risk assessment.

2. Pre-specify "risk thresholds" that limit the level of risk accepted

Describe and continually refine risk assessment results for each model ("risk thresholds") that would trigger particular risk-reducing actions, defining such results in terms of risk to all relevant stakeholders given currently existing mitigations. Given the high uncertainty around future model capabilities, risk thresholds may be refined periodically.

Define risk thresholds, based on the outcomes that would constitute a breach of the threshold and linked to dangerous capabilities that a given model or combination of models could exhibit. For example, a frontier AI organisation might identify the objective of avoiding deploying an AI system that significantly increases the risk of cyberattacks or fraud.

Operationalise risk thresholds, including specific, testable observations, such that multiple observers with access to the same information would agree on whether a given threshold had been met. Specific observations would provide frontier AI organisations with opportunities to determine proactively how they would respond in difficult potential situations, and so respond immediately to such situations should they arise, as well as allowing for accountability and external verification.

Given the nascent science of AI evaluation, however, it is unlikely to be possible to define a set of testable observations that detect all identified risks sufficiently reliably. Instead of relying solely on these predefined tests, risk assessments may take into account wider sources of evidence, such as concerning and unexpected observations that show up in exploratory analyses, expert forecasts, or risk related information from other frontier AI organisations. In particular, consideration may be given to any risks caused by the combination of a given model with other models or tools, whether developed by the same frontier AI organisation or otherwise.

Continue to refine and redefine risk evaluation frameworks for models as necessary, aiming to reduce the gap between the intended objectives of risk thresholds and their present operationalisations. Such gaps are expected to exist due to limitations in the science of evaluation and in the state of knowledge surrounding capabilities, so progress towards a robust framework will probably be iterative. Risk evaluation frameworks may use multiple methods, including probability estimations and qualitative assessments of current capabilities.

Mitigate the risk of 'overshooting' thresholds. This may be achieved by setting deliberately conservative thresholds, including using intentionally lower buffer thresholds to trigger actions, such that the most concerning thresholds are difficult to overshoot without having already implemented mitigations at an earlier stage.

Engage with relevant external stakeholders when developing risk thresholds. Risk thresholds often concern externalities frontier AI organisations place on society, including both the potentially significant benefits of AI advancement and negative effects that might disproportionately affect specific stakeholder groups. As such, their risk thresholds may be made public to allow for external scrutiny, with thresholds set in consultation with relevant external stakeholders including relevant government authorities.

3. Pre-commit to specific additional mitigations at each risk threshold, followed by a residual risk assessment

At each risk threshold, proactively commit to only proceed with certain development or deployment steps if specific mitigations are in place. Such mitigations could include many of the practices outlined in this document.

After putting in place mitigations, reassess any residual risk posed to determine whether additional mitigations are required. Due to the unpredictability of capabilities advancements and the limitations of model evaluation science, pre-agreed mitigations may prove insufficient to place a given model within a risk threshold.

Risk acceptance criteria may be used, as is standard in many other contexts, and may provide an important tool for clarification. These criteria may evolve with time, and could be quantitative or qualitative. For example, risk may only be accepted if it has been reduced to a level 'as low as reasonably practicable'.

Inform relevant government authorities when a risk threshold has been met, along with proposed mitigations. Inform governments again, in advance of deployment, when the mitigations and residual risk assessment have been carried out. Proactively engage other relevant actors in addition to governments as appropriate.

4. Adapt mitigations to the relevant stages of development and deployment, recognising that models may be used in different ways and contexts than intended

When planning required mitigations, consider the full range of development and deployment stages. In general, meeting a risk threshold could require mitigations at multiple such stages. Important stages may include the following:

- Continued training of model
- Deployment of model in small-scale ways, e.g., internal use
- Deployment of model in large-scale ways e.g., public release via API
- Extension of model through greater affordances, e.g., tool use or internet access
- Irreversible deployment e.g. open-sourcing of models

Adapt mitigations in recognition of risks from unintended model use or use in unexpected contexts, such as a model that is modified to remove safeguards after open-sourcing or a model that is combined with another model for unanticipated purposes. For example, at a given risk threshold, information security control mitigations might be put in place before even internal use of a model within a frontier AI organisation. In general, the use of caution may be helpful, given current limitations in information security controls and the prediction of models' emergent abilities.

In particular, recognise that even the possession of some models may be dangerous, even if not deployed or not deployed widely, due to the risk of being unable to secure a model sufficiently to prevent, for instance, a bad actor obtaining the model weights. In contrast, other models may pose significant risk only if deployed in a large-scale or irreversible way.

Deploy models in small-scale or reversible ways before deploying models in large-scale or irreversible ways. This makes it possible for frontier AI organisations to notice and mitigate harm before the harm becomes too large or unavoidable.

5. Make preparations to pause development and/or deployment should risk thresholds be reached without the pre-agreed mitigations

Prepare to pause training runs or reduce access to deployed models, if risk thresholds are reached without the committed risk mitigations being in place. This may involve warning existing customers that access reductions are a possibility and creating contingency plans to minimise negative impacts on customer use.

Consider the potential risks of open-sourcing models. Acknowledge that some models may pose additional risks if made available “open-source”, even after mitigation attempts. This is because of the inability of recalling an open-sourced model and the potential ability of users to remove safeguards and introduce new (and potentially dangerous) capabilities. However, it is also important to bear in mind the significant benefits of “open-source” AI systems for researchers, including for advancing AI safety, which may in some cases outweigh these potential risks.

6. Share details of risk assessment process and risk mitigation measures with relevant government authorities and other AI companies

Regularly update relevant stakeholders on risk assessment and mitigation measures: This will enable assessment of whether AI organisations have sufficient risk management processes in place, build up a picture of best practices, and make recommendations to address gaps. When sharing this information with external actors, consideration should be given to commercially sensitive information. Additional ad hoc updates could be provided in cases of major developments.

Include information on evaluations, risk assessment and mitigation, and individuals involved. For example:

- What types of tests and evaluations are being run on which types of models
- What other risk assessment methods are being used, which kinds of expertise are drawn on, and whether impacted stakeholders are being involved
- How risk mitigation measures are being monitored
- Which teams and individuals are involved at different stages of the risk management process (and how, if at all, third parties are involved)
- Measures taken to address specific categories of risk, such as cybersecurity measures

Consider publicising high-level summaries of risk assessment and mitigation processes to enable wider scrutiny and build public confidence in the safety and reliability of AI systems. Sensitive details could be removed.

7. Commit to robust internal accountability and governance mechanisms, alongside external verification

Introduce robust and meaningful accountability mechanisms, especially in evaluating capabilities thresholds, with clear processes that ensure the correct mitigations or courses of action are followed if the thresholds are met. This may include board sign-off for the responsible capability scaling policy, and named individual accountability for key decisions.

Establish effective risk governance to ensure that risks are appropriately identified, assessed, and addressed, and their nature and scale transparently reported. Most importantly, provide internal checks and balances, which may include thoughtful separation of roles within risk management.

Include verification mechanisms, such that external actors can have increased confidence that responsible capability scaling policies are executed as intended.

Potential mechanisms for information sharing are included in Model Reporting and Information Sharing.

Model Evaluations and Red Teaming

Summary

Frontier AI may pose increased risks of harm related to misuse, loss of control, and other societal risks. Different methods are being developed to assess AI systems and their potential harmful impacts. Model evaluations- such as benchmarking- can be used to produce quantitative, easily replicable measurements of the capabilities and other traits of AI systems. Red teaming provides an alternative approach, which involves observing an AI system from the perspective of an adversary to understand how they could compromise or misuse it.

Assessments like model evaluations and red teaming could help to understand the risks frontier AI systems pose and their potential harmful impacts, and help frontier AI organisations, regulators, and users to make informed decisions about training, securing, and deploying them. As methods for assessing frontier AI systems are still emerging, it is important to support and share information about the development and testing of these methods.

We outline four categories of practice regarding Model Evaluations and Red Teaming:

1. **Evaluate models for several sources of risk and potential harmful impacts**, including dangerous capabilities, lack of controllability, societal harms, and system security

2. **Conduct model evaluations and red teaming at several checkpoints throughout the lifecycle of a model**, including during and after training and fine-tuning, as well as post-deployment

3. **Allow independent, external evaluators to conduct model evaluations throughout the model lifecycle, especially pre-deployment**

4. **Support advancements in the science of model evaluation**

Background

Understanding the capabilities and limitations of frontier AI systems is crucial to their effective governance. This forms the basis of risk assessments and, ultimately, responsible development and deployment. Sharing this understanding, where appropriate and safe to do so, can also provide important transparency to external actors.

Gaining an understanding of system capabilities and limitations is challenging. Often, these are only discovered after a model has been deployed, used by millions of users, and integrated into downstream products.

The purpose of model evaluations and red teaming is to help generate this understanding to inform the responsible development, deployment and use of frontier AI systems. By investing more into finding out the relevant information both before and after these models are deployed, developers and society at large can understand the capabilities and limitations of these models sooner.

Independent, external evaluation can help to verify developers' claims around the safety of their frontier AI systems. Although the third party audit market is currently nascent,[1] this will grow as more organisations implement the practice.

Practices

1. Evaluate models for several sources of risk and potential harmful impacts, including dangerous capabilities, lack of controllability, societal harms, and system security

Evaluate models for potential dangerous capabilities (i.e. capabilities that could cause substantial harm either from intentional misuse or accident). These capabilities could include but are not limited to:

- Offensive cyber capabilities (e.g. producing code to exploit software vulnerabilities)
- Deception and manipulation (e.g. lying effectively or convincing people to take costly actions)
- Capabilities that can assist users in developing, designing, acquiring, or using biological, chemical, or radiological weapons (e.g. helping users "troubleshoot" their efforts to produce biological weapons)

Evaluate models for controllability issues[2] (i.e. propensities to apply their capabilities in ways that neither the models' users nor the models' developers want). This could include, for example, autonomous replication and adaptation (i.e. capabilities that could allow a model to copy and run itself on other computer systems).

Evaluate models for societal harms. These could include, for example, bias and discrimination (e.g. the risk that they produce content that reinforces harmful stereotypes or their potential discriminatory influence if used to inform decisions), recognising that 'bias' can be difficult to define and can be subject to different interpretations in different contexts.

Evaluate models for system security (see Security Controls Including Securing Model Weights section).

Ensure processes are in place to respond to evaluation results. Evaluations are a necessary input to a responsible capability scaling policy which, depending on the results of the evaluation, would probably require implementation of practices outlined in other sections of this document such as preventing model misuse, information sharing and other risk mitigation measures.

2. Conduct and support red teaming and model evaluations at several checkpoints throughout the lifecycle of a model, including during and after training and fine-tuning, as well as post-deployment

Before a frontier model is trained, evaluate predecessor or analogous models to understand how relevant properties (e.g. dangerous capabilities) scale with the overall size of the model. These preliminary evaluations can inform risk assessments.

[1] Digital Regulation Cooperation Forum, 'Auditing Algorithms'.
[2] Controllability is also sometimes referred to as "alignment" or "steerability".

During pre-training and fine-tuning, evaluate the model to detect signs of undesirable properties and identify inaccuracies in pretraining predictions. These evaluations could be undertaken at various pre-specified checkpoints, and could inform decisions about whether to pause or adjust the training process.

After training, subject the model to extensive pre-deployment evaluations. These evaluations can inform decisions about whether and how to deploy the system, as well as allowing governments and potential users to make informed decisions about regulating or using the model. Their intensity will be proportional to the risk of the deployment, taking into account the model's capabilities, novelty, expected domains of use, and number of individuals expected to be affected by it.

After deployment, conduct evaluations at regular intervals to identify new capabilities and associated risks, especially when notable developments (e.g. a major update to the model) suggest earlier evaluations have become obsolete. Post-deployment evaluations can inform decisions to update the system's safeguards, increase security around the model, temporarily limit access, or roll back deployment.

Require organisations who deploy their models to conduct context-specific model evaluations. This requires that the information and data required to successfully conduct such assessments is provided to deployers.

3. Allow independent, external evaluators to conduct model evaluations, especially pre-deployment

Evaluation by independent third parties will allow frontier AI organisations to draw on external expertise, have more "eyes on the problem", and provide for more accountability. External evaluation is particularly important at the pre-deployment phase, and can inform irreversible decisions around deployment of the model. Appropriate legal advice and confidentiality agreements may also protect any market-sensitive data when sharing information with third parties. For the subset of evaluations which may touch on national security concerns, a secure environment may be needed with appropriately cleared officials. There are further opportunities for independent evaluation for open source models, given the potential broader community involvement.

Ensure that evaluators are independent and have sufficient AI and subject matter expertise across a wide range of relevant subjects and backgrounds. External evaluators' relationships with frontier AI organisations could be structured to minimise conflicts of interest and encourage independence of judgement as far as practically possible. As well as expertise in AI, there are many other areas of subject matter expertise that will be needed to evaluate an AI system's features. For instance, experts on topics as wide as fairness, psychological harm, and catastrophic risk will be needed.

Ensure that there are appropriate safeguards against external evaluations leading to unintended widespread distribution of models. Allowing external evaluators to download models onto their own hardware increases the chance of the models being stolen or leaked. Therefore, unless adequate security against widespread model distribution can be assured, external evaluators could only be allowed to access models through interfaces that prevent exfiltration (such as current API access methods). It may be appropriate to limit evaluators' access to information that could indirectly facilitate widespread model distribution in other ways, such as requiring in-depth KYC checks or watermarking copies of the model.

Give external evaluators sufficient time. As expected risks from models increase or models get more complex to evaluate, the time afforded for evaluation may need to increase as well.

Give external evaluators the ability to securely "fine-tune" the AI systems being tested. Evaluators cannot fully assess risks associated with widespread model distribution if they cannot fine-tune the model. This may involve providing external evaluators with access to capable infrastructure to enable fine-tuning.

Give external evaluators access to versions of the model that lack safety mitigations. Where possible, sharing these versions of a model gives evaluators insight into the risks that might be created if users find ways to circumvent safeguards (i.e. "jailbreak" the model). If the model is open-sourced, leaked, or stolen, users may also simply be able to remove or bypass the safety mitigations.

Give external evaluators access to model families and internal metrics. Frontier AI organisations often develop "model families" where multiple models differ along only one or two dimensions – such as parameters, data, or training compute. Evaluating such a model family would enable scaling analysis to better forecast future performance, capabilities and risks.

Give external evaluators the ability to study all of the components of deployed systems, where possible. Deployed AI systems typically combine a core model with smaller models and other software components, including moderation filters, user interfaces to incentivise particular user behaviour, and plug-ins for extension capabilities like web browsing or code execution. For example, a red team cannot find all the flaws in the defences of a system if they aren't able to test all of its different components. It is important to consider the need to balance external evaluators' ability to access all components of the system against the need to protect information that would allow bypassing model defences.

Allow evaluators to share and discuss the results of their evaluations, with potential restrictions where necessary e.g. not sharing proprietary information, information whose spread could lead to substantial harm or information that would have an adverse effect on competition in the market. Sharing the results of evaluations can allow governments, regulators, users, and other frontier AI organisations to make informed decisions.

4. Support advancements in the science of model evaluation

Support the development and testing of model evaluation methods. For many relevant properties of models, there do not yet exist accepted evaluation methods. It also remains unclear how reliable or predictive current evaluation methods are. This could involve frontier AI organisations developing model evaluation methods themselves or facilitating the efforts of others, such as by providing access to capable infrastructure for evaluation.

Share the products of their model evaluation research and development, except when sharing the results might be harmful. In some cases, findings (e.g. about how to elicit dangerous capabilities) could be harmful if spread. When the expected harm is sufficiently small, the AI research community, other frontier AI organisations, and relevant government bodies could benefit from being informed of their work.

Model Reporting and Information Sharing

Summary

Transparency around frontier AI can help governments to effectively realise the benefits of AI and mitigate AI risks. Transparency can also encourage sharing of best practices across frontier AI organisations, enable users to make well-informed choices about whether and how to use AI systems, and increase public trust, helping to drive AI adoption.

Reporting and sharing information where appropriate could ensure that different parties can access the information they need to support effective governance, develop best practice, inform decision-making about the use of AI systems, and build public trust. Some reporting practices- such as model cards- are already used among frontier AI organisations, whereas other practices included here are areas for future consideration.

Given the recent rapid pace of progress in AI, the appropriate government and international governance institutions are still being considered and we recognise that limits the ability of frontier AI organisations to share information with governments, even where it would be desirable. Throughout this section “relevant government authorities” is used to indicate a good practice for information sharing with governments while recognising such relevant authorities may still be under development.

We outline three categories of practice regarding Model Reporting and Information Sharing:

1. **Share model-agnostic information about general risk assessment, mitigation and management processes and best practices**

2. **Share model-specific information about certain frontier AI models** before training, during training, and before deployment

3. **Share different information with different parties**, including government bodies, other frontier AI organisations, independent third parties, users, and the public, as appropriate

Background

There is currently a large information asymmetry between those developing frontier AI systems and those who oversee and make use of them, which is compounded by the pace of AI progress. Improving access to information on how AI systems are developed can help overseers hold frontier AI organisations accountable for safe and responsible development of AI and enable users of these systems to effectively manage their risk.

Standardised documentation for consumer products is commonplace, e.g. information leaflets for medications and nutritional information on food packaging. Transparency reports—often in the form of model cards—have accompanied a number of recent major model releases, and good practices for transparency reporting are beginning to emerge.

Public transparency reports are clearly distinguishable from other information-sharing channels, such as between frontier AI organisations and governments. Some information that is shared with governments or regulators is not necessarily appropriate to be shared publicly. For example, it may be good for frontier AI organisations to report summary results of model evaluations for dangerous capabilities, whereas publicly reporting detailed techniques for

eliciting such dangerous capabilities may be harmful. To protect market sensitive technical capabilities, commercially sensitive information could be shared with government or regulators, who could share this information with industry in an aggregated or anonymised form.

Practices

1. Share model-agnostic information about general risk assessment, mitigation and management processes and best practices

Share details of risk assessment processes and risk mitigation measures with relevant government authorities and other AI companies, as set out in Responsible Capability Scaling.

Share information about how internal governance processes are set up with relevant government authorities. This will ensure that risks are appropriately identified, communicated and mitigated, and allow government and other external actors to identify gaps that might lead to risks being overlooked. This information could be updated regularly (e.g. every 12 months). This information could also be made public, provided sensitive details are removed.

Report any details of security or safety incidents or near-misses to relevant government authorities. This includes any compromise to the security of the organisation or its systems, or any incident where an AI system – deployed or not – causes substantial harm or is close to doing so. This will enable government authorities to build a clear picture of when safety and security incidents occur and make it easier to anticipate and mitigate future risks. Incident reports could include a description of the incident, the location, start and end date, details of any parties affected and harms occurred, any specific models involved, any relevant parties responsible for managing and responding to the incident, as well as ways in which the incident could have been avoided. It is important that incidents indicative of more severe risks are reported as soon as possible after they occur. High-level details of safety and security incidents - with sensitive information removed - could also be made public, such as have been shared in the AI incident database.

2. Share model-specific information about certain frontier AI models before training, during training, and before deployment

Sharing information about specific frontier AI models allows external actors to develop a more granular picture of ongoing AI development and potential risks that will need to be addressed.

Before training, share high-level model details with the relevant government authorities and justify why the training run does not impose unacceptable risk.

This could include:

- A high-level description of the model (including high-level information on intended use cases, intended users, training data, and model architecture)
- Compute details (including the maximum the organisation plans to use, as well as information about its location and who provides it)
- Description of the data that will be used to train the model
- Evidence from scaling small models that the full training run does not pose unacceptably high risks

- Descriptions of specific internal and external risk assessments and mitigation efforts,[3] and an overall safety assessment justifying why and how the training run is sufficiently low-risk to execute
- Description of which, if any, domain experts and impacted stakeholders have been involved in the project's design, as well as risk and impact assessment
- Plans for model evaluations during and after training, as well as predicted dangerous capabilities

During training, update information provided to relevant government authorities with any significant changes to the model itself or its risk profile.

This could include:

- Updates on model development at each evaluation checkpoint as well as any significant updates to the development plan
- Results from model evaluations, including details of emergent dangerous capabilities and whether these were expected
- Whether and how the risk context has changed (e.g. if other AI tools have been published that the model could interact with)

At the point of deployment, share details of the model, any risks the model might pose and what steps have been taken to mitigate those risks with relevant government authorities and the wider public.

Information can be provided in full to the relevant government authority, ensuring that robust security measures are in place to protect sensitive information. Some of this information can be made available to the public by publishing a transparency report (e.g. a model card) and providing general overviews of model purpose and risk assessment evaluation results. Information that exposes model vulnerabilities or facilitates the spread of dangerous capabilities should be redacted from the transparency report, unless sharing this information publicly would be sufficiently helpful for mitigating the risks the model poses. Information that is commercially sensitive (e.g. detailed information on training data) or exposes the capabilities or vulnerabilities of the model should be redacted. Commercially sensitive information could be shared with government or regulators, who could share this information with industry in an aggregated or anonymised form.

Information shared at the point of deployment could include, for example:

- Details about the model, such as:
  - A description of the model
  - Information about safe practices for model usage (including domains of inappropriate and appropriate use, or guidelines on determining whether a use is appropriate)
  - Training details, including a detailed description of the training data and any biases they may encode
  - The model's biases, risks and limitations

[3] See Responsible Capability Scaling section.

- Information about risk and impact assessments, governance mechanisms, and capability evaluations, such as:
  - Pre-development and pre-deployment risk and impact assessment procedures
  - Details of the evaluation process the frontier AI organisation conducted, including time and resources spent, information about the expertise and independence of people conducting the evaluations, the level of access given to evaluators, and anticipated limitations of the evaluations used
  - Details about which, if any, domain experts and impacted stakeholders have been involved in the project's design and risk and impact assessment
  - The results of any internal or external evaluations that were conducted
  - Holistic assessments of the models' robustness, interpretability, and controllability, drawing on more specific evaluation results
  - Significant measures taken to mitigate potential harmful consequences of deployment, including accountability and verification mechanisms put in place, and internal governance processes carried out
  - Capabilities and risks of the final, public-release model before and after safety mitigations, including a description of the mitigations to prevent accidents and misuse (e.g. available tools and their expected effectiveness)
  - Plans for ongoing, post-deployment monitoring of risks and capabilities and how the organisation will respond to future incidents (unless releasing this information would allow bad actors to circumvent post-deployment safety measures)
- Other information about the model, such as:
  - Descriptions of post-deployment access controls
  - Expected compute requirements for running the model during deployment

3. Share different information with different parties, including relevant government authorities, other frontier AI organisations, independent third parties, and the public, as appropriate

Before sharing information, consider the risks of sharing this information and judge whether it is inappropriate to share certain pieces of information. In particular, it is important to consider potential harms from publicly sharing information about dangerous capabilities and methods for eliciting them, as this information could motivate or help other actors to acquire these capabilities. It is also important to consider potential harms from publicly sharing detailed information about how models were produced, as this may lower barriers to producing similar models. If the models have or could be modified to possess dangerous capabilities (e.g. biological capabilities or surveillance capabilities), then facilitating widespread distribution of the model in this way may be harmful. The National Protective Security Authority (NPSA) guidance on a security-minded approach to information management may prove helpful for sharing information appropriately.

Develop principled policies about what information to share publicly, with governments, or not at all. These policies could specify situations under which sharing some piece of information is subject to a risk assessment, along with guidelines for conducting the risk assessment and responding to it. It is important, however, to avoid creating risk assessment criteria and procedures that are overly strict and intensive to prevent excessive opacity. These policies could be guided and overseen by an independent review panel of multidisciplinary

experts to ensure decisions made about or against information sharing are justifiable and oriented to optimal transparency.

Share complete forms of this information with central AI-focused government authorities (including central government bodies, security agencies, and regulators) to enable robust government oversight of potentially high-risk areas of AI development and the processes in place to identify and manage those risks. These authorities could then further share information selectively with additional government bodies where relevant. Some particularly security-relevant pieces of information may need to be shared directly with security agencies, such as the cybersecurity measures being used in model development (which would make it easier for those agencies to identify potential security risks), or specific physical or cybersecurity threat incidents. Robust security measures are in place when sharing more sensitive information.

Share more limited information with other AI organisations in order to facilitate learning and the development of best practices. This could include best practices and lessons learned in the development of risk assessment and mitigation measures and risk governance processes, some details about safety and security incidents (to enable increased awareness while protecting intellectual property), and highlights and lessons learned from risk assessments and capability evaluation. In general, it will be easier for organisations to share model-agnostic information with one another than model-specific information, which may be more commercially sensitive. It will also be important to consider that there is no privileged access to information that may give a competitive advantage to some firms.

Share specific information with independent third-parties where this aids evaluation and technical audit (see Model Evaluations and Red Teaming). This may require sharing much of the same information that is shared with governments in full but may be shared on a case-by-case basis.

Share specific information with downstream users of the model to enable more effective risk mitigation across the AI supply chain and build consumer confidence. This could include uses of the model that are against their terms of service, and could be provided to users through user terms or published information about the model.

Share more general versions of the information outlined with the public to enable public scrutiny and build public confidence in the safety and reliability of AI systems. This could include high-level summaries of risk assessment and mitigation processes, high-level details of safety and security incidents, summaries of risk governance processes, and general overviews of model purpose, use cases, and the results of risk assessments and capability evaluations.

Security Controls Including Securing Model Weights

Summary

To ensure the safety of frontier AI, consideration of cyber security, protective security risk management and insider risk mitigation is key. Cyber security, both of models and the systems that deploy them, must be considered from the outset of development to ensure that the benefits of AI can be realised. Cyber security is a key underpinning for the safety, reliability, predictability, ethics and potential regulatory compliance of an AI system.

To avoid putting safety or sensitive data at risk, it is important to consider the cyber security of AI systems, as well as models in isolation, and to implement cyber security processes throughout the AI lifecycle, particularly where that component is a foundation for other systems. As AI systems advance, developers must maintain an awareness of possible attacks, identify vulnerabilities and implement mitigations. Failure to do so will risk designing vulnerabilities into future AI models and systems. A Secure by Design approach allows developers to 'bake in' security from the outset of design and development.

Cyber security must be considered in concert with physical and personnel security. Developing a coherent, holistic, risk based and proportionate security strategy, supported by effective governance structures, is essential. Where the compromise of an AI system could lead to tangible or widespread physical damage, significant loss of business operations, leakage of sensitive or confidential information, reputational damage and/or legal challenge, then it is important that AI security risks are treated as mission critical.

We outline eight categories of practice with regards to Security controls, Including Securing Model Weights:

1. **Implement strong cyber security measures and processes (including security evaluations) across their AI systems, including their underlying infrastructure and supply chains**

2. **Understand the assets in their AI system (including training data and model weights) and take appropriate action to protect them**

3. **Maintain current understanding of cyber security risk, including novel threats to and from AI systems, to enable informed risk decisions**

4. **Include consideration of AI in incident response, escalation and remediation plans and ensure responders have been trained to assess and address AI-related incidents**

5. **Perform ongoing monitoring of system behaviour so that changes in behaviour may be observed and potential attacks identified**

6. **Enable users' secure use of AI systems by evaluating and communicating risks and following Secure by Design principles**

7. **Implement effective protective security risk management – covering physical, personnel and cyber security disciplines**

8. Develop and implement appropriate personnel security controls to mitigate insider risk

Background

As the use of AI and machine learning (ML) systems continues to grow, systems must be developed and deployed securely to avoid putting safety or data at risk. It is important to consider the cyber security of AI models and systems and to implement cyber security processes throughout the AI lifecycle, monitoring security on a continuous basis.

The National Cyber Security Centre (NCSC) has published a range of guidance to help organisations boost their resilience and protect themselves online. Existing cyber security good practice provides a strong foundation for AI security. However, there are additional measures that can help to address inherent security weaknesses in the way AI systems and ML models work. In addition to standard cyber security failure modes, such as exploitation of traditional vulnerabilities in underlying software stacks, it is important to consider the potential for AI-specific failures.

Cyber security must be considered alongside physical and personnel security, for which the NPSA has published guidance to help protect organisations from a range of threats. High profile supply chain compromises, such as the SolarWinds breach, demonstrate why it is crucial to quickly discover and mitigate vulnerabilities in ubiquitous software. AI is no exception, and due to the complex nature of many supply chains, it can be difficult to know if there is enough protection across the supply chain. The NPSA has produced Protected Procurement guidance to help different audiences understand and address supply chain security and the NCSC has produced guidance specifically aimed at helping organisations to assess the cyber security of their supply chain.

Recruitment of insiders is an attractive option for threat actors to access valuable material, and implementing measures to reduce the risk of insider activity and to protect the organisation's assets is important. Resources for understanding and managing insider risk are published by the NPSA.

Practices

1. Implement strong cyber security measures and processes (including security evaluations) across their AI systems, including underlying infrastructure and supply chains

Apply good infrastructure principles to every part of this process from design to decommissioning. This is important as threats can occur at different stages of an AI project's life cycle.

Regularly assess the security of their supply chains and ensure suppliers adhere to the same standards their own organisation applies. Ensuring data, software components and hardware are obtained from trusted sources will help mitigate supply chain risk.

Evaluate the robustness of models to different classes of adversarial attack (such as poisoning, model inversion and model stealing), based on priorities derived from threat modelling. This could involve a combination of benchmarking and red teaming.

Assess and document the cyber security threats against the AI system overall and mitigate the impact of vulnerabilities. Good documentation and monitoring will inform your overall risk posture and help you respond in the event of a security incident.

2. Understand the assets in their AI system and take appropriate action to protect them

Understand the value of AI-related assets such as models, data (including user feedback), prompts, software, documentation, and assessments (including information about potentially unsafe capabilities and failure modes) to their organisation. Protect these different categories of information as appropriate.

Ensure security is factored into all business decisions and AI-related assets are identified and protected with proportionate cyber, physical and personnel security measures. Secure Innovation guidance from NCSC and NPSA is available to help companies and investors to protect their technology .

Have processes and tools to track, authenticate, secure and version control assets and be able to roll back to a known safe state in the event of a compromise. This is important as data and models do not remain static during the whole lifespan of an AI project.

Document data, models, prompts, evaluation materials and other assets, using commonly used structures such as data cards, model cards and software bills of materials (SBOMs). This will allow you to identify and share key information, including particular security concerns, quickly and easily.

3. Ensure developers and system owners maintain current understanding of security risk to enable informed risk decisions

Train developers, system owners and senior leaders in secure AI practices. It is crucial to establish a positive security culture where leaders demonstrate good security practice and staff from across a project understand enough about AI security to understand the potential consequences of decisions they take.

Maintain an awareness of security threats and failure modes, in particular data scientists and developers. AI development and cyber security are two different skill sets and building a team at the intersection of the two will require effort and time.

Model the threats to your system to understand the impacts to the system, users, organisation and wider society if the model is misused or behaves unexpectedly. This can help build systems where unanticipated model outputs are handled safely by other parts of a data pipeline.

4. Develop incident response, escalation and remediation plans and ensure responders have been trained to assess and address AI-related incidents

Develop an organisational incident response plan. The inevitability of security incidents affecting systems is reflected in organisational incident response planning. A well-planned response will help minimise the damage caused by an attack and support recovery.

5. Perform ongoing monitoring of system behaviour so they can observe changes in behaviour and identify potential attacks

Measure the performance of AI models and systems on an ongoing basis. A decline in model performance could be an indication of an attack or could indicate that a model is encountering data that is different from that which it was trained on. Either way, further investigation may be required.

Monitor and log inputs to AI systems to enable audit, investigation and remediation in the case of compromise. Some attacks against AI systems rely on repeated querying. Proper logging will help you audit your system and identify any anomalous inputs.

Take action to mitigate and remediate issues and document any AI-related security incidents and vulnerabilities.

6. Enable secure use of AI systems by users, by communicating risks and following Secure by Design principles

Communicate clearly to users which elements of security they are responsible for and where and how their data may be used or accessed.

Integrate the most secure settings within your products by default. Where configuration is necessary, default options should be broadly secure against common threats.

Ensure necessary security updates are a default part of every product and use secure, modular update procedures to distribute them. This will help your products remain secure in the face of new and developing threats.

Only release models after they have been through security evaluations, including benchmarking and red teaming.

Maintain open lines of communication for feedback regarding product security, both internally and externally to your organisation, including mechanisms for security researchers to report vulnerabilities and receive legal safe harbour for doing so, and for escalating issues to the wider community. Helping to share knowledge and threat information will strengthen the overall community's ability to respond to AI security threats.

7. Implement effective protective security risk management – covering physical, personnel and cyber security disciplines

Together with defined governance and oversight, key steps towards effective protective security risk management include, but are not limited to:

- Identify assets & systems that are important for the delivery of effective operations, or are of specific organisational value (e.g. commercially sensitive information).
- Categorise and classify assets in order to ensure that the correct level of resource is used in implementing risk mitigations.
- Identify threats. These may include terrorism or hostile state threats and/or more local and specific threats, and use a range of internal and external resources.
- Assess risk using recognised processes.

- Build a protective security risk register to record, in sufficient detail, all the data gathered during this risk management process, ensuring compatibility with existing organisational risk management registers and processes.
- Develop a protective security strategy for mitigating the risks identified, which reviews protective security measures in relation to a prioritised list of risks. Where mitigations are assessed as inadequate, additional measures could be proposed for approval by the decision maker(s).
- Produce development & implementation plans. Aim to arrive at a clear, prioritised list of protective security mitigations, which span physical, personnel and cyber security disciplines, and are linked to the technical guidance needed to implement them.
- Review risk management measures regularly and when required e.g. on a change in threat or change to operational environment, or to assess the suitability of new measures implemented. More detailed description of protective security risk management is provided on the NPSA website.

8. Develop and implement appropriate personnel security controls to mitigate insider risk

Key steps towards mitigation of insider risk include but are not limited to:

- Ensure board-level responsibility for protective security with regular engagement with key stakeholders from across the business and a firm understanding of the risks the organisation faces. Ensure stakeholder engagement throughout the business for specialist insight and development and implementation of an insider risk mitigation programme.
- Apply a suitable level of screening, informed by a role-based risk assessment, to all individuals who are provided access to organisational assets including permanent, temporary and contract workers.
- Use Role-Based Security Risk Assessment to identify physical, personnel or cyber security measures that need to be applied in order to mitigate insider risk
- Put in place proportionate policies, clear reporting procedures and escalation guidelines that are accessible, understood and consistently enforced.
- Provide appropriate security education and training for all workers. Without effective education and training individuals cannot be expected to know what procedures are in place to maintain security.
- Ensure that a programme of monitoring and review is in place to enable potential security issues, or personal issues that may impact on an employee's work, to be recognised and dealt with effectively throughout their career
- Use established, evidence based guidance to fully address personnel security risks e.g. NPSA guidance on Personnel Security

Reporting Structure for Vulnerabilities

Summary

Even after a frontier AI organisation has deployed an AI system, the system may still have unidentified safety and security issues ("vulnerabilities"). In order to address these vulnerabilities, they must first be identified, and frontier AI organisations should be made aware of them.

Establishing a vulnerability management process enables outsiders to identify and report any vulnerabilities. This can help to ensure that safety and security issues are flagged to frontier AI organisations as soon as possible so they are able to address them quickly.

We outline three categories of practice regarding Vulnerability Reporting Structures:

1. **Establish a vulnerability management process**
2. **Establish clear, user-friendly, and publicly described processes for receiving model vulnerability reports** drawing on established software vulnerability reporting processes
3. **Develop protocols and mechanisms for coordinated vulnerability disclosure and information sharing**

Background

AI models can have unexpected vulnerabilities. For example, though developers continually patch them, users have been repeatedly able to perform jailbreaks of large language models, using specific prompts to make them behave in ways counter to developer intentions, which can be harmful. The importance of addressing such vulnerabilities will increase in tandem with model capabilities. Should such issues remain and model capabilities be sufficiently dangerous, significant restrictions on model use may be warranted.

One way to improve developers' ability to address vulnerabilities is to incentivise users and researchers outside the frontier AI organisation to help address the issue, as a complement to robust evaluation and risk assessment processes. Well-established vulnerability reporting protocols will often pair notification of the relevant organisation with public disclosure after some delay. That way, the affected organisation is incentivised to address the vulnerability.

Given the differences between the conventional software vulnerabilities and AI model vulnerabilities, the bug bounty model cannot be straightforwardly applied to model vulnerabilities. Compared to conventional software vulnerabilities, it can be unclear how to go about fixing model vulnerabilities once they are discovered, which may make public disclosure of vulnerabilities less appropriate. Secondly, compared to traditional software vulnerabilities, it may be especially difficult to specify what qualifies as a model vulnerability ahead of time.

Practices

1. Establish a vulnerability management process[4]

This process could have as wide a scope as is necessary and ensure that frontier AI organisations have the ability to respond appropriately to reports of vulnerabilities. The process could accept reports on any class of model vulnerabilities and methods for exploiting them, including:

- **Jailbreaking methods**: methods for inducing models to bypass moderation features, which the frontier AI organisation has attempted to prevent through the use of filters or fine-tuning
- **Prompt injection attacks**: methods used by malicious actors to induce models to exhibit behaviours they want by presenting models with prompts that contain instructions to perform these behaviours
- **Privacy attacks**: methods for extracting information that should be private from models (e.g. sensitive information from training data or users' private conversations with models)
- **Unaddressed misuse opportunities**: methods for using the capabilities of models to cause harm, which have not already been addressed
- **Controllability issues (i.e. "misalignment")**: when models apply their capabilities in ways that substantially diverge from what users intend or desire
- **Poisoning attacks**: when an adversary has manipulated training data in order to degrade model performance
- **Bias and discrimination**: when a model is exhibiting behaviours that reveal specific biases or discrimination regarding known protected characteristics
- **Performance issues**: when a model performs inadequately for a situation it is being deployed in e.g. a healthcare chatbot AI that provides incorrect information and causes harm to patients

2. Establish clear, user-friendly, and publicly described processes for receiving model vulnerability reports drawing on established software vulnerability reporting processes

These processes can be built into – or take inspiration from – processes that organisations have built to receive reports of traditional software vulnerabilities. It is crucial that these policies are made publicly accessible and function effectively.

3. Develop protocols and mechanisms for coordinated vulnerability disclosure and information sharing

Consider the ways in which sharing information about vulnerabilities can both exacerbate and mitigate risks. Sharing can alert would-be attackers, but also alert would-be victims and actors with the power to create defences. One particularly important factor is how

[4] See NCSC's Vulnerability Disclosure Toolkit for guidance on vulnerability disclosure.

easily *fixable* the model vulnerability is. If a vulnerability will take a very long time to fix, or cannot be fixed, then public reporting may not be justified.

Developing – and publicly describe – protocols for deciding how to share model vulnerability information. These protocols may, for example, outline conditions under which information is to be shared with different actors depending on the type of harm or vulnerability identified.

Put in place mechanisms to disclose information about vulnerabilities to relevant government authorities, law enforcement, and other affected organisations. This may be particularly relevant for vulnerabilities where public disclosure might increase the risk of harm.

Publicly share general lessons learned from model vulnerability reporting programs. This might include, for example, lessons about challenges faced and the relative efficacy of different incentive strategies. Such sharing could be done via a mechanism similar to the National Institute for Science and Technology's National Vulnerability Database in the US.

Identifiers of AI-Generated Material

Summary

Some content generated by AI can be difficult to distinguish from content generated by a human. Where harmful or false information is generated and spread by bad actors using AI, this can pose risks to public safety. The ability to distinguish between human and AI-generated content is therefore increasingly important.

AI identifiers can aid the identification of AI-generated content but can be technically challenging to implement in practice and currently cannot entirely mitigate risks from well-resourced actors. While authentication solutions (e.g. 'watermarking') are under development, they may not be considered fully reliable, as there are techniques that may allow users to escape detection.

Adopting a range of identifier mechanisms and investing in the development of techniques that allow AI-generated content to be identified could mitigate a variety of risks including (but not limited to) those associated with creating and distributing deceptive AI-generated content, content bias in AI generation and loss of trust in information.

We outline three categories of practice with respect to Identifiers of AI-Generated Material:

1. **Research techniques that allow AI-generated content to be identified**
2. **Explore the use of watermarks for AI-generated content, including those that are robust to various perturbations**
3. **Explore the use of AI output databases**

Background

Some AI-generated content can appear indistinguishable from human-generated content. This raises several risks, including bad actors creating and spreading harmful or false information. A number of actors- including individuals, companies, institutions and governments- will become increasingly reliant on the ability to distinguish AI generated content from human generated content.

AI identifiers are technical measures which can aid in differentiating AI-generated content from non-AI generated content. These include watermarking, and database-based solutions such as those that record outputs at the point of generation. While AI identifiers are limited and cannot fully mitigate AI risks, their widespread adoption could mean the most commonly used AI generation services have a degree of friction for harmful uses of AI content, reducing the incentives to pass AI generated content off as real.

However, there are potential challenges with some of these techniques. Some watermarking techniques can be easily circumvented, particularly by actors of moderate to high sophistication. There are also privacy and security challenges around creating databases of content known to be AI generated, although hashing technologies - which create a hash (usually a string of characters) that is unique to content at the point it is generated - have the potential to mitigate some of these issues. Looking at the current state of the art, no specific identifier is impossible to circumvent or guaranteed to be technically feasible at scale.

However, collective implementation of multiple identifier mechanisms could add friction and cost to creating and distributing deceptive AI-generated content, and may mitigate an unintentional AI-driven decline of information quality.

Practices

1. Research techniques that allow AI-generated content to be identified

Invest in researching how AI-generated content may be watermarked, including AI-generated text, photos and videos, and experiment with the implementation of such techniques. It is particularly technically difficult to attach watermarks and prove the provenance of text. Researching how this content can be watermarked, including experimenting with the implementation of techniques, may help with these challenges. One method could involve making the model more statistically likely to use certain phrases or words in a way that is unnoticeable to humans, but can be picked up by a detector, provided a long enough sequence of text. However, this approach may not be robust to attempts to scrub off the watermark e.g. by having another AI model paraphrase the text.

2. Explore the use of watermarks for AI generated content that are robust to various perturbations

Explore the use of watermarks for AI generated content that are robust to various perturbations after their creation, including attempts at removal. To make the removal of watermarks more difficult, developers of generative AI models may need to consider how they distribute certain information about their watermarking methods or open-sourcing their classifiers. This may also include monitoring ways in which adversarial users are attempting to scrub off their watermarks and patching, where relevant, such means of circumvention. It also includes a recognition that watermarking however may not be appropriate in all circumstances given the limitations.

3. Explore the use of AI output databases

Explore databases of content generated or manipulated by a model to identify AI-generated content. These databases could be queried by third parties, including auditors and regulators, facilitating identification of potentially AI-generated content. Such databases could include only a subset of generated content, which is flagged as potentially important. To ensure user privacy, privacy-preserving techniques could be explored in conjunction with such databases, such as hashing technologies. This allows for the identification of AI-generated content without the need to store the actual content, thereby respecting user privacy. Additionally, common standards between different databases from various frontier AI organisations could facilitate a unified search, allowing for the identification of AI-generated content across all frontier AI organisations simultaneously.

Prioritising Research on Risks Posed by AI

Summary

The future capabilities and risks associated with frontier AI are uncertain, and continued research is required to better understand them. Frontier AI organisations are uniquely positioned to conduct research on the risks that frontier AI poses and develop tools to address them. As gatekeepers to information that is critical for research on AI risks, frontier AI organisations have a significant role to play in facilitating open and robust research across the AI ecosystem.

We outline four categories of practice for Research on Risks Posed by AI, including:

1. **Conduct research to advance AI safety**
2. **Invest in developing tools for defending against harms and risks from their systems** (e.g. watermarking tools to defend against misinformation)
3. **Collaborate with external researchers to study and assess the potential downstream social impacts of their systems** (e.g. on employment and the spread of misinformation)
4. **Publicly share the products of their risk research** (except when sharing these products might cause harm)

Background

AI is a rapidly evolving field and increasingly capable and complex models continue to be developed and released, and the AI ‘frontier’ will continue to shift. To identify and mitigate these risks, continued research will be required.

Frontier AI organisations have an important role to play in this research ecosystem, as they have access to critical AI inputs that can be directly put towards mitigation (e.g. compute, data, talent, and technical know-how). There are also measures that frontier AI organisations are uniquely well-positioned to take, such as using proprietary models to create defensive tools or making their models less prone to misuse or accidents.

However, addressing the harmful consequences of frontier models will require close and wide collaboration between frontier AI organisations, other AI organisations and external actors. Frontier AI organisations need to consider the sensitivity of the research they are conducting and the potential for theft, misuse or exploitation.

Practices

1. Conduct research to advance AI safety

Conduct research, collaborating with external stakeholders and organisations as needed, to identify and mitigate the risks and limitations of AI, including research on:

- **Interpretability and explainability**: improving our ability to understand the inner functioning of AI systems and explain their behaviour
- **Evaluation**: improving our ability to assess the capabilities, limitations, and safety-relevant features of AI systems
- **Robustness**: improving the resilience of AI systems e.g. against attacks intended to disrupt their proper functioning
- **Reliability and controllability (or "alignment")**: improving the consistency of an AI system in adhering to the specifications it was programmed to carry out and operating in accordance with the designer's intentions, and decreasing its potential to behave in ways its user or developer does not want (e.g. producing offensive or biased responses, not refusing harmful requests, or employing harmful capabilities without prompting)
- **Bias and discrimination**: improving our ability to address bias and discrimination in AI systems
- **Privacy**: improving our ability to address the privacy risks associated with AI systems
- **Hallucinations**: decreasing AI systems' (specifically large-language models') propensity to generate false information
- **Cybersecurity**: improving our ability to ensure the security of AI systems
- **Criminality**: improving our ability to prevent criminal behaviour through the use of AI systems (e.g. fraud, online child sexual abuse)
- **Other societal harms**: improving our ability to prevent other societal harms arising from the use of AI systems, including psychological harm, misinformation, and other societal harms.

2. Invest in developing tools for defending against harms from their systems

When a probable and consequential harm from a frontier AI organisation's system is identified, investigate whether there are tools that can be built to mitigate this harm. For instance, recognizing the rise in AI-generated child exploitation and abuse content, some social media platforms are developing tools for identifying and removing child sexual abuse content.

Work closely with external actors who will need to deploy the tools to ensure that they are usable and suit their needs. For instance collaborating closely with social media organisations to help them produce more capable tools for identifying AI-generated content.

Make special efforts to ensure that the defensive tools are available at or before the time of system release. The larger the risk is and the more effective tools may be, the more important it is to prepare defensive tools ahead of time. It may be important to delay system releases until appropriate defensive tools are ready.

Disseminate defensive tools responsibly, sometimes sharing them publicly and sometimes sharing them only with particular actors. In some cases, making a tool freely available (e.g. by open-sourcing it) may reduce its effectiveness by allowing malicious actors to study it and circumvent it.

Continuously update defensive tools as workarounds are discovered. In some cases, this may be an ongoing effort that requires sustained investment.

3. Collaborate with external researchers to study and assess the potential downstream societal impacts of their systems

Study the societal impacts of AI systems they have deployed, particularly by collaborating with external researchers, independent research organisations, and third party data owners. By collaborating and joining their data with that of third parties, such as internet platforms, frontier AI organisations can assess the impacts of their AI systems. Privacy-enhancing technologies could be employed to enable data sharing between frontier AI organisations, third parties, and external researchers while protecting confidential information. As well as data, frontier AI organisations could also facilitate research on the societal impacts of their AI systems by providing access to the necessary infrastructure and compute.

Draw on multidisciplinary expertise and the lived experience of impacted communities to assess the downstream societal impacts of their AI systems. Impact assessments that account for a wide range of potential societal impacts and meaningfully involve affected stakeholder groups could help to anticipate further downstream societal impacts.

Use assessments of downstream societal impacts to inform and corroborate risk assessments. Downstream societal impacts, such as threats to democracy, widespread unemployment, and environmental impacts, could be considered in risk assessments of AI systems, alongside more direct risks. See the Responsible Capability Scaling section for more information on good practices for risk assessment.

Ensure equitable access to frontier AI systems. Transparent and fair processes for researchers to get restricted access to AI systems are important. To ensure systems are appropriately understood, particular attention could be paid to promote academic freedom and diversity of thought, for example, not withholding access based on previous or expected criticism and encouraging different types of academics, civil society groups and independent researchers to study AI systems.

4. Publicly share the products of their risks research, except when sharing these products might cause harm

In the absence of sufficiently substantial downsides to sharing, frontier AI organisations are encouraged to share the products of this work broadly.

Preventing and Monitoring Model Misuse

Summary

Once deployed, AI systems can be intentionally used to achieve harmful and/or illegal outcomes. These include cyberattacks, spreading disinformation, and criminal activity.

To prevent model misuse, it is important to monitor and respond to misuse of their systems. This can help to ensure that instances of misuse are identified and dealt with promptly, mitigating the risk of more widespread harm.

We outline six categories of practice regarding Preventing and Monitoring Model Misuse:

1. **Set up processes to identify and monitor misuse of models** such as by monitoring for common ways models are misused and safeguards are circumvented
2. **Implement model input and output filters**
3. **Implement additional measures to prevent harmful outputs** including fine-tuning, prompting, and rejection sampling
4. **Implement user-based API access restrictions and monitoring** e.g. reducing access to individuals who repeatedly trigger content filters without reasonable justification
5. **Prepare to respond to potential worst-case or consistent misuse scenarios** including via rapid model rollback and withdrawal
6. **Continually assess the effectiveness and desirability of existing and additional safeguards** since they may also hinder positive uses and reduce privacy

Background

AI systems can be misused in many ways, including for cyberattacks (e.g. spear phishing) and spreading disinformation. AI systems can also increase the scale and speed of criminal offending, and are increasingly being used for crime such as fraud, online child sexual abuse, and intimate image abuse. While the potential for future misuse can be reduced during model training, steps can also be taken to identify and respond to such misuse during model deployment.

Frontier systems are often deployed via API services, through which access to the model is controlled by the AI organisation and users can query the model without running it on their own hardware, as opposed to publicly releasing model weights (sometimes referred to as “open-source” or “open-access” models). This section focuses on models deployed via API services.

Open-source access to low-risk models may enable greater understanding of AI system safety through greater public scrutiny and testing of AI systems. At the same time, developers of open source models typically have less oversight of downstream uses, meaning that many of the methods for preventing the misuse of higher-risk models addressed in this section are unavailable to frontier AI organisations releasing open-source models. Such organisations can still identify and mitigate model misuse to some extent e.g. by endeavouring to enforce open-source model licences. However, releasing models via APIs provides significantly more

affordances for frontier AI organisations to address misuse as they maintain visibility on how models are being used, increased control over usage and safeguards, and the ability to update or roll back a model after deployment.

It is important to continuously evaluate and adjust practices to prevent and mitigate model misuse as usage changes. This includes considering which forms of use warrant prevention, since the distinction between misuse and legitimate use can be ambiguous.

Practices

1. Set up processes to identify and respond to patterns of misuse of models

Understand how their own models – as well as how models released by other groups – are being misused. Knowing that some people have begun using a competitor's model to conduct phishing attacks, for example, may lead a frontier AI organisation to become more concerned that its own model will be misused in this way.

Report information on broad patterns of misuse. This information can help governments, the public, and other frontier AI organisations to better understand risks. Reports may focus on population-level metrics related to API usage, such as the rate of harmful inputs filtered or the number of users banned for misuse. This information may benefit from being presented in a clear and understandable way and made optimally accessible.

Determine appropriate retention schedules for usage logs, balancing safety and privacy considerations. In severe cases of misuse, access to logs from several months prior may be necessary to understand in maximal detail the causes of the misuse. However, in some cases extended retention schedules may disproportionately affect privacy.

2. Implement model input and output filters

Apply content filters to both model inputs and model outputs. The input filters can block harmful requests (e.g. requests for advice on building weapons) from being processed by the model. Content modifiers could adjust harmful prompts to elicit non-harmful responses. The output filters can block harmful model outputs (e.g. instructions on building weapons) from being sent back to the user.

Explore and compare the efficacy of multiple approaches to developing content filters. This may involve comparing available methods, such as those in the following section; selecting the most effective ones; and applying them in combination if doing so substantially increases efficacy. Best practice could be shared with other frontier AI organisations or published broadly.

Invest in making content filters robust to "jailbreaking" attempts. Work to ensure filters are robust to jailbreaking attempts, for instance by including examples of jailbreaking attempts in the datasets used to develop filters.

Exercise appropriate caution when developing, storing, or sharing content filters – and any specialised datasets used to produce them. In some cases, components of filters or datasets used to create them can be used to train higher-risk models. For instance, a classifier that evaluates the aggressiveness of a model's outputs may be used to train the model to be more aggressive.

3. Implement additional measures to prevent harmful outputs

Fine-tune models to reduce the tendency to produce harmful outputs. This may involve using reinforcement learning from human or AI-generated feedback regarding the appropriateness of different model outputs e.g. a constitutional based approach where human input to the fine-tuning process is provided by a list of principles. It may also involve fine-tuning models on curated datasets of appropriate responses to prompts. Where human feedback is used, the mental wellbeing of moderators may need to be considered.

Prompt models to avoid harmful behaviour. This may involve, for instance, using a "meta prompt" to instruct a model that it should ignore requests to cause harm. Alternatively, prompt distillation or other methods can be used to fine-tune their models to behave as though they have received particular prompts.

Consider also applying "rejection sampling" methods to model outputs. These methods involve generating several outputs, scoring them on their harmfulness, and then only presenting the least harmful outputs to the user.

4. Implement user-based API access restrictions

When deploying models through APIs, consider reducing access to users who display suspicious usage patterns. For instance, if a content filter blocks a user's requests several times in short succession, this is evidence that they are attempting to misuse the model or find ways to circumvent its filters. Appropriate measures may include warnings, further investigation, rate limitations, more restrictive filters, and bans. Care should be taken to avoid restricting genuine use, for example to legitimate AI safety researchers attempting to examine model behaviour, or to users struggling to access legitimate sources of help on difficult topics like self-harm.

Communicate API access reduction policies – and allow users to appeal access reductions. Users should be able to understand the reasons why their access may be reduced and be able to appeal access reductions if they believe policies have not been applied correctly. Alternatively, users may be able to perform actions to mitigate misuse risks, such as providing evidence to justify behaviour.

Implement tiered Know Your Customer (KYC) checks for API users. KYC checks can help prevent users from simply creating new accounts when their access is reduced. More intense checks, such as identify verification, could be implemented where the risk is higher, such as for models with more dangerous capabilities, access to models with fewer safeguards, or high-volume usage of the model. It is important to weigh up KYC checks against potential privacy and access tradeoffs of requiring registrations.

Consider restricting certain API access tiers only to users in "trusted" categories. For example, it may be appropriate to offer high usage rates, fine-tuning access, permissive content filters, as well as access to models with high misuse potential only to verified users in established enterprises, non-profits, and universities.

Establish protocols for deciding when and how to share information about API users with relevant government authorities. These protocols may include covering circumstances under which information about a user is proactively shared with government bodies (e.g. cases where there is reason to think a user may be attempting to cause grave harm), and how to respond to government requests for data.

5. Prepare to respond to potential worst-case or consistent misuse scenarios, including via rapid model rollback or withdrawal

Implement processes and technical requirements to enable rapid rollback or withdrawal of models in case of egregious, widespread or consistent harms. Rolling back to a previous version of a model that does not suffer from the same misuse threats, or withdrawing access to a model altogether, may be proportionate actions in such situations of extreme threat or consistent misuse. Running periodic dry runs could increase preparedness for such situations.

Inform end users that rapid rollback or withdrawal of models may be necessary, and that the disruption to its end users will be minimised as far as safely possible. This is especially important where models are deployed in safety critical domains.

Establish governance processes to ensure internal clarity in response to worst-case or consistent misuse scenarios. Such processes could draw on similar accountability and governance mechanisms used within a Responsible Capability Scaling policy. It may be valuable to clarify the approvals needed to roll back a model to a previous version or withdraw a model, the types of misuse that would warrant such actions, and the timeframe under which such actions would occur.

6. Continually assess the effectiveness and desirability of existing and additional safeguards

Regularly assess monitoring and safeguard efficacy and continually invest in improvements. Regular assessment of monitoring efficacy can build an understanding of the success rate and speed of issue detection, on which deployment decisions may be based. These assessments may draw, for instance, on the results of internal and external red-teaming efforts, random audits of usage logs, what may be ‘best practice’, as well as available information about real-world misuse. Risks are expected to increase as AI capabilities increase. Hence, active increases in investment in safety, security, and monitoring may be valuable.

Explore new safeguards and countermeasures in response to patterns of misuse, recognising that appropriate measures may vary based on model type, usage patterns, and our technical understanding of model capabilities.

Employ diverse monitoring techniques to balance comprehensiveness, scalability, and privacy. A strong monitoring setup will generally combine automated and human review, in recognition that automated reviews may miss intricate issues, while human reviews are not always scalable and can pose privacy issues.

Regularly assess the costs of safeguards to users, including bias and loss of privacy, as well as restrictions on users’ freedom to discuss sensitive topics with AI agents in appropriate contexts. This may involve, for instance, estimating the false positive rates of filters by sampling blocked inputs and outputs; comparing the capabilities, efficiency, and user ratings of models that lack certain safeguards against models that have the safeguards in place; and conducting user interviews to understand how concerned informed users are about losses in privacy and the emergence of biases in algorithmically filtered content.

Explore strategies for reducing the costs of safeguards to users, such as “structured transparency” methods to reduce the cost of usage monitoring to user privacy, or tiered access to models with different safeguards for differently qualified users.

Data Input Controls and Audits

Summary

The data used to train AI systems influences how they behave. Where frontier AI systems are trained on poor quality or undesirable data, this increases the risks they pose and could enhance their potential dangerous capabilities.

By controlling and auditing the data AI systems are trained or fine-tuned on, it is possible to make more accurate predictions about their capabilities and mitigate risks by, for example, removing input data that may produce an AI system with dangerous capabilities. Data input controls and audits can also provide important information to downstream users and regulators.

We outline four categories of practice with respect to Data Input Controls and Audits. These are:

1. **Before collecting training data, implement responsible data collection practices**
2. **Audit input data before it is used to train the AI system** e.g. attempting to identify data that could yield dangerous capabilities
3. **Put in place appropriate risk mitigations in response to data audit results** e.g. by curating their datasets to ensure they are not training on certain data
4. **Facilitate external scrutiny of input data** by inviting external actors to assess their input data and sharing information from their input data audits

Background

The data that AI systems are trained on influences how they behave and the risks they pose. Though the relationship between training data and system behaviour is complex, some predictions are possible.

Data that is biased, inaccurate, or inadequately representative may lead AI systems to perform worse and have more detrimental societal impacts. AI systems that are trained on sensitive or personal data may be vulnerable to extraction attacks. Some training data may enhance AI systems' potential dangerous capabilities, making them more harmful when misused.[5]

These relationships are problematic for frontier AI systems, which are often trained on large datasets which may contain such data. Moreover, frontier AI systems are further modified by additional data via fine-tuning or reinforcement learning from human feedback (RLHF). These smaller supplementary datasets exert a disproportionately large influence on the performance and behaviour of AI systems.

Responsible data collection practices can help to make AI systems safer upfront by improving the quality of training data, thereby reducing the likelihood of systems being trained on dangerous, sensitive, or imbalanced data. Careful audits of input data can also yield better

[5] Markus Anderljung et al, 'Frontier AI Regulation: Managing Emerging Risks to Public Safety'.

predictions about system capabilities and limitations, while revealing avenues to reducing risk e.g. by not training on certain undesirable data.

Practices

1. Before collecting input data, implement responsible data collection practices

Before collecting training data, take account of applicable regulatory frameworks. This may involve, for example, establishing a legal basis to process training data and understanding any copyright considerations that might apply. This could help to mitigate risks further down the line, such as the system revealing personally identifiable information.

Consider the principle of data minimisation where possible. Data minimisation can reduce the risk of harmful content making it into training data. Frontier AI organisations could explore the practice of "data pruning", which has been shown to improve data quality and system performance while minimising the quantity of pre-training data required.

2. Audit input data before it is used to train the AI system

Audit datasets used for pre-training but also those used for fine-tuning, classifiers, and other tools. Inappropriate datasets could result in systems that fail to disobey harmful instructions.

Use technical tools – such as classifiers and filters – to audit large datasets to support scalability and privacy. These could be used in combination with human oversight, which can verify and augment these assessments.

Assess the overall composition of training data. This could include the data sources, the provenance of the data, indicators of data quality and integrity, and measures of bias and representativeness. The amount and variety of data are simple, reliable predictors of risk, and provide an additional line of defence where more targeted assessments are limited.

Audit datasets for:

- **Information that might enhance dangerous system capabilities**, such as information about weapons manufacturing or terrorism.
- **Private or sensitive information**. AI systems may be subject to data extraction attacks, where determined users can prompt systems to reveal pieces of training data, or may even reveal this information accidentally. This makes it important to know whether datasets include private or sensitive information e.g. names, addresses, or security vulnerabilities.
- **Biases in the data**. Training data that is imbalanced or inaccurate can result in an AI system being less accurate for people with certain personal characteristics or providing a skewed picture of particular groups. Ensuring a better balance in the training data could help to address this.[6]

[6] See Centre for Data Ethics and Innovation, 'Review into bias in algorithmic decision-making'.

- **Harmful content**, such as child sexual abuse materials, hate speech, or online abuse. Having a better understanding of harmful content in datasets can inform safety measures (e.g. by highlighting domains where additional safeguards like content filters should be applied).
- **Misinformation**. Training an AI system on inaccurate information increases the likelihood the outputs of the system will be inaccurate and could lead to harm.

Draw on external expertise in conducting input data audits. For example, biosecurity experts could be consulted to identify information relevant to biological weapons manufacturing, which may not be readily obvious to non-experts.

Use data audits to improve understanding of how training data affects AI system behaviour. For example, if model evaluations reveal a potentially dangerous capability, data audits can help ascertain the extent to which the training data contributed to it.

Conduct audits on datasets used by their customers to fine-tune AI systems. Customers are often allowed to fine-tune systems on their own datasets. By carrying out audits to ensure that customers are not encouraging undesirable behaviours, frontier AI organisations can use their expertise and insight into the AI system's original training data to identify potential harms upstream. It is important that frontier AI organisations are mindful of privacy concerns and make use of privacy preserving techniques, where appropriate.

Document the results of input data audits, including metadata. Frontier AI organisations could look to emerging standards when documenting the results of input data audits, such as datasheets for datasets.[7]

3. Put in place appropriate risk mitigations in response to data audit results

Use input data audits to inform pre-development and pre-deployment risk assessments. Data audits may be particularly valuable in pre-development risk assessments, where direct evidence of system behaviour will not be available. Input data audits can also be used to improve context-based understanding of how data impacts system capabilities, which will strengthen future risk assessments and mitigation techniques.

Remove potentially harmful or undesirable input data, where appropriate. Given the increasingly strong generalisation abilities of AI systems, data curation may prove insufficient to prevent dangerous system behaviour but could provide an additional layer of defence alongside other measures such as fine-tuning and content filters.

For some kinds of risks, explore options to use harmful data to reduce AI systems' dangerous capabilities or to help develop mitigation tools. For example, by fine-tuning a system using labelled harmful content to refuse requests related to the harmful information.

[7] See Timnit Gebru, et al, 'Datasheets for Datasets', and Emily Bender and Batya Friedman, 'Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science' for proposed frameworks for dataset reporting.

Consider sourcing or generating additional data and adding it to the training dataset, where it is determined that data is missing or inadequate. Improving the representativeness of training data can improve performance and reduce potential negative societal impacts and discriminatory effects. However, additional data should only be sought through appropriate means that respect and empower those individuals who are missing from the data.

Acknowledge the limitations of input data audit. Techniques for understanding the impacts of data and filtering out specific data are limited (e.g. excluding information from a dataset does not always prevent the AI system from reasoning about or discovering that information). Other risk mitigation measures will be required and further research on improving data input audit techniques is important.

4. Facilitate external scrutiny of training datasets

Facilitate independent data input audits from external parties. Since training data constitutes sensitive intellectual property, it is important to implement appropriate technical and organisational safeguards to ensure privacy and security when sharing the training data. In order to protect sensitive information, frontier AI organisations could explore the possibility of providing synthetic datasets to auditors with sensitive information removed, or providing auditors with privacy-preserving access to the training data.

Share information about input data audits with users and external stakeholders. Some information could be included in transparency reports, such as high-level information about training data (including data sources), data auditing procedures, and measures taken to reduce risk (see Model Reporting and Information Sharing). More sensitive information may be shared directly with regulators and external auditors.

Be mindful that techniques for identifying dangerous information may be susceptible to misuse. For example, techniques for identifying private information in large datasets could be used by cybercriminals, and should therefore only be shared responsibly. Dangerous or otherwise sensitive information identified during audits should be stored securely to avoid leaks.

Acknowledgements

We thank the following for their contributions to this document, including providing comments on drafts, discussing research questions, and suggesting practices:

Ada Lovelace Institute
Alan Turing Institute
Alignment Research Center
British Computer Society
Center for Security and Emerging Technology, Georgetown University
Centre for the Governance of AI
Centre for Long Term Resilience
Professor John Tasioulas, Sir Nigel Shadbolt, Dr Caroline Green, and Professor Jeremias Adams-Prassl (all of the University of Oxford Institute for Ethics in AI)
Future of Life Institute
RAND Corporation

We also shared the document with the frontier AI companies publishing safety policies ahead of publication.