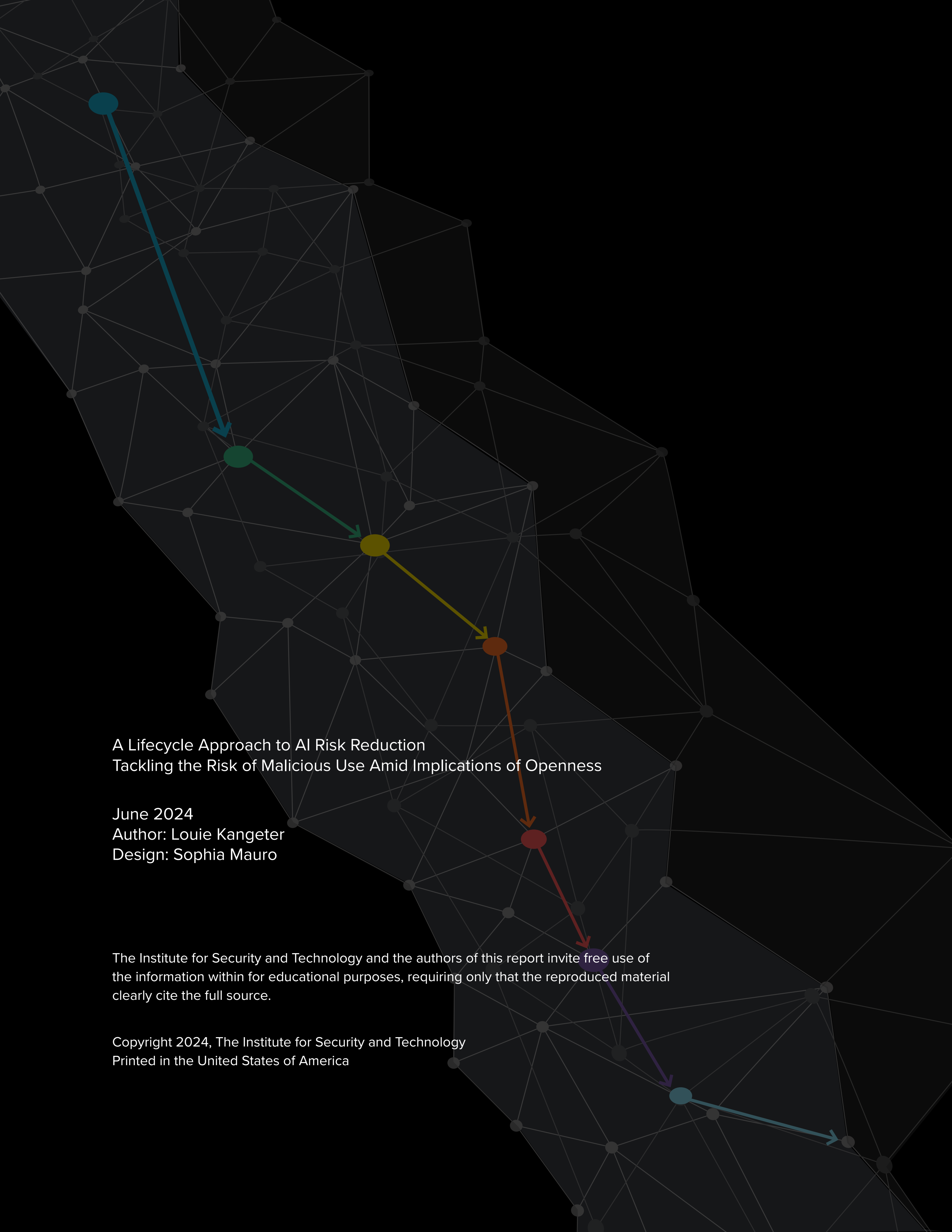
A LIFECYCLE APPROACH TO AI RISK REDUCTION

TACKLING THE RISK OF MALICIOUS USE AMID IMPLICATIONS OF OPENNESS

LOUIE KANGETER

JUNE 2024

IST Institute for SECURITY + TECHNOLOGY



A Lifecycle Approach to AI Risk Reduction
Tackling the Risk of Malicious Use Amid Implications of Openness

June 2024
Author: Louie Kangeter
Design: Sophia Mauro

About the Institute for Security and Technology

The Institute for Security and Technology (IST) is a global 501(c)(3) think tank uniquely situated in Silicon Valley with deep ties to Washington, D.C. and other global capitals.

As new technologies present humanity with unprecedented capabilities, they can also pose unimagined risks to global security. IST's mission is to bridge gaps between technology and policy leaders to collaboratively solve these emerging security challenges *together*. We have the access and relationships to unite the right experts via bespoke convening mechanisms, with agility, and at the right time. We are translators, conveners, and communicators who leverage unique problem-solving approaches to tackle some of the world's toughest emerging security threats.

Our portfolio is organized across three analytical pillars: the **Geopolitics of Technology**, anticipating the positive and negative security effects of emerging, disruptive technologies on the international balance of power, within states, and between governments and industries; **Innovation and Catastrophic Risk**, providing deep technical and analytical expertise on technology-derived existential threats to society; and the **Future of Digital Security**, examining the systemic security opportunities and risks of societal dependence on digital technologies.

Acknowledgments

This work is inherently collaborative. As researchers, conveners, and facilitators, the Institute for Security and Technology (IST) is immensely grateful to the members of this working group for their insights, dedication, willingness to engage in honest and healthy debate, and the time that each of them generously volunteered to this effort. By working collectively to identify where risk is arising and where it is likely to continue to intensify, their contributions are what allowed this process to bear fruit.

We are also immensely grateful for the generous support of the **Patrick J. McGovern Foundation**, whose funding allowed us to continue this project through the lens of IST's Applied Trust & Safety program, and **Google.org**, whose funding supports IST's AI in cybersecurity work.

AI is too vast a set of tools, capabilities, and communities for any one organization to manage the risks and opportunities on its own. This effort reflects the cross-sectoral, public-private efforts needed more broadly across the ecosystem to ensure AI is beneficial for us all. While each working group member does not necessarily endorse everything written in this report, we extend our gratitude to the following contributors and editors in particular:

- Chloe Autio
- Zoë Brammer
- Vaishnavi J
- Jeffrey Ladish
- Mariami Tkeshelashvili

Additionally, not everyone in the working group could choose to be named openly as contributors. We are just as grateful to them, including those individuals we have worked closely with and who helped to inspire the initial effort.

Finally, the author extends his gratitude to Steve Kelly and Philip Reiner for the support, guidance, and strategy they provided in the drafting and refining of this report.

Contents

Executive Summary

The rapid advancement and proliferation of artificial intelligence (AI) technologies has brought forth myriad opportunities and challenges, necessitating the development of comprehensive risk mitigation strategies. Building on IST's December 2023 report, How Does Access Impact Risk? Assessing AI Foundation Model Risk Along a Gradient of Access—which evaluated six categories of AI risk across seven levels of model access—this report provides policymakers and regulators with a robust framework for addressing these complex risks.

The report establishes five guiding principles that serve as the foundation for the proposed risk mitigation strategies: balancing innovation and risk aversion, fostering shared responsibility among stakeholders, maintaining a commitment to accuracy, developing practicable regulation, and creating adaptable and continuous oversight.

Central to the report is the AI Lifecycle Framework, which builds on working group contributions, mainly the "upstream/downstream" framing of risks and mitigations, and breaks down the complex process of AI development into seven distinct stages: data collection and preprocessing, model architecture, model training and evaluation, model deployment, model application, user interaction, and ongoing monitoring and maintenance. By identifying the most effective points for implementing risk mitigations within each stage, the framework enables targeted interventions that align with the guiding principles.

To demonstrate the application of the AI Lifecycle Framework, the report conducts a deep dive into malicious use—one of the risks identified in the December 2023 report as negatively influenced by an increased gradient of model openness—examining five key areas: fraud and crime schemes, the undermining of social cohesion and democratic processes, human rights abuses, disruption of critical infrastructure, and state conflict. The analysis considers the historical context, current state-of-play, and outlook associated with each area.

Applying the AI Lifecycle Framework to malicious use risks reveals a range of effective risk mitigation strategies at each stage of the AI lifecycle. These strategies encompass both policy and technical interventions, such as introducing incentives for ethical data collection practices, developing secure model architectures, and implementing human oversight in high-risk AI applications. Additionally, the report acknowledges the limitations and challenges of risk mitigation throughout the gradient of open access models and emphasizes the need for ongoing research, collaboration, and adaptation.

The report concludes with a call for continued exploration of risk mitigation strategies across other risk categories and collaboration among stakeholders to refine and implement the proposed strategies. By proactively addressing AI risks while fostering innovation, the AI Lifecycle Framework serves as a valuable tool for guiding effective risk mitigation efforts in the face of rapid technological advancements.

Introduction

Following our December 2023 report, *How Does Access Impact Risk? Assessing AI Foundation Model Risk Along a Gradient of Access* (hereafter, "the Phase I report")[1]—which evaluated six categories of AI risk across seven levels of model access—IST identified a need to provide policymakers and regulators in the global community with strategies for mitigating these risks. Through our conversations with experts in the field, we consistently received feedback emphasizing the importance of equipping policymakers and regulators with a conceptual understanding of the AI landscape and, more crucially, the overarching goals that inform risk mitigation strategies. To provide this context, this report establishes a set of guiding principles, enabling readers to contextualize AI risk mitigation strategies within a goals-based framework.

The report then introduces an AI Lifecycle framework, demonstrating how understanding the AI lifecycle and consciously building risk mitigations to fit appropriately within it allows for the mitigations themselves to be most effective. Additionally, the report takes a deep dive into one of the six previously identified risk categories—malicious use—examining past and present malicious behaviors and presenting possible routes bad actors may take in the future. This sets the stage for targeted risk mitigation approaches.

The report is structured as follows: the Guiding Principles section outlines the guiding principles that serve as the foundation for our risk mitigation strategies. The AI Lifecycle Framework section introduces the AI Lifecycle Framework, which helps identify the most effective points for implementing risk mitigations and connecting the framework to our guiding principles. While this framework is applicable to all AI models, from open access models to proprietary ones, this report pays special attention to one of the risks negatively influenced by model openness. The Deep Dive on Malicious Use provides a comprehensive analysis of past, present, and potential future threats. The section entitled Applying the Framework presents specific risk mitigation strategies based on the insights gained from the previous sections.

[1] Zoë Brammer, "How Does Access Impact Risk? Assessing AI Foundation Model Risk Along a Gradient of Access," Institute for Security and Technology, December 2023, https://securityandtechnology.org/wp-content/uploads/2023/12/How-Does-Access-Impact-Risk-Assessing-AI-Foundation-Model-Risk-Along-A-Gradient-of-Access-Dec-2023.pdf.

Mapping Openness Over the Lifecycle provides crucial context on the interaction of model access with risk mitigation strategies. Finally, the Conclusion offers recommendations for future policy and regulation.

Utilizing the AI Lifecycle framework, and a deep understanding of malicious behaviors, this report applies knowledge gleaned from working group sessions and IST research to identify several key risk mitigations that are most likely to be effective in reducing or preventing malicious use of AI systems. By providing this comprehensive analysis and actionable recommendations, it aims to support policymakers and regulators in developing informed strategies to address the complex challenges posed by AI technologies.

While this report addresses only one of the six risk categories identified in our Phase I report, this approach, and indeed this report, can be extended in future phases of work to address them.

The Guiding Principles

These guiding principles are the result of extensive discussions with working group members, who emphasized the importance of creating a flexible and adaptable framework that can accommodate the diverse and evolving nature of AI risks. The guiding principles presented in this section are designed to provide high-level context within which policymakers, AI developers, and other stakeholders can navigate the complex landscape of AI risk mitigation. By adhering to these principles, this report aims to foster a balanced, collaborative, and proactive approach to addressing the challenges posed by AI systems, ensuring that the benefits of this transformative technology are maximized while potential harms are minimized.

PRINCIPLE #1: BALANCING INNOVATION AND RISK AVERSION

Mitigating AI risks requires balancing innovation and caution. This principle emphasizes creating an environment that encourages responsible innovation while prioritizing the identification, assessment, and mitigation of potential risks. Achieving this balance will enable society to harness AI's power to drive progress while ensuring its development and use align with safety, ethics, and trustworthiness.

PRINCIPLE #2: SHARED RESPONSIBILITY

Effective AI risk mitigation demands collaboration from all stakeholders, including policymakers, AI developers, users, and civil society. Each group contributes unique perspectives, expertise, and roles, while introducing potentially risky behavior or outcomes. Recognizing and embracing this shared responsibility fosters a collaborative approach to risk mitigation, leveraging the strengths and contributions of all stakeholders.

PRINCIPLE #3: COMMITMENT TO ACCURACY

Ensuring that AI models provide reliable and factual information is crucial for building trust and fostering the beneficial application of this technology across various domains. Risk mitigation strategies should prioritize the preservation of accuracy and avoid resorting to censorship, factual alteration, or the compromising of truth for the sake of more agreeable outcomes. By upholding this principle, policymakers can create a regulatory framework that encourages the development of AI systems that are not only trustworthy but also genuinely useful, as their outputs can be relied upon to inform critical decisions and ultimately shape humanity's understanding of complex challenges.

PRINCIPLE #4: PRACTICABLE REGULATION

Developing regulatory oversight for AI requires considering the technical feasibility of proposed measures. Practicable oversight involves collaboration with AI experts and stakeholders to identify and implement mechanisms that are both effective and technically feasible. This may necessitate exploring alternative approaches that provide meaningful accountability without imposing technically infeasible barriers or constraints to AI development.

PRINCIPLE #5: ADAPTABLE AND CONTINUOUS OVERSIGHT

Regulatory frameworks should keep pace with the latest technological advancements, best practices, and lessons learned. These frameworks should incorporate mechanisms for regular review and refinement, as well as continuous monitoring and oversight. This approach involves establishing data collection, analysis, and feedback loops to inform iterative improvements and risk mitigation efforts, while maintaining vigilance towards emerging threats, vulnerabilities, and ethical concerns.

Guiding Principles in Action

These principles create a balanced and comprehensive approach to AI risk mitigation. IST recognizes the importance of fostering innovation while prioritizing safety and ethics, emphasize shared responsibility and collaboration among stakeholders, and underscore the importance of adaptable, feasible, and continuously evolving regulatory frameworks and oversight mechanisms. Embracing these principles ensures that risk mitigation efforts are effective, sustainable, and aligned with the dynamic nature of AI technologies and their societal implications. Armed with these principles, this next section explores structuring risk mitigation strategies across each phase of the AI lifecycle.

AI Lifecycle Framework

In January 2024, IST hosted a working group meeting to kick off the second study phase of the AI Foundation Model Access Initiative, an effort to assess the risks and opportunities of increased access to AI foundation models. The working group coalesced around the idea of identifying and examining both risks and mitigations that are "upstream" and "downstream" across the AI lifecycle. Note, there exists a distinction between upstream/downstream *risks* and upstream/downstream *mitigations*.

- **Upstream risks**: Risks inherent to a model itself which result from the model's training and development. Examples include fueling a race to the bottom and capability overhang.[2,3]
- **Downstream risks**: Risks that result from user interactions with models. Examples include malicious use and taking the human out of the loop.[4,5]
- **Upstream mitigations**: Mitigations that target model development and pre-deployment training.
- **Downstream mitigations**: Mitigations that target model release, fine tuning by third parties, the development of applications built on foundation models, and user interaction.

This distinction is important because risk *mitigation* can occur both upstream and downstream across the AI lifecycle, even for categories of risk that seemingly fall in the opposite category. For example, while malicious use is a downstream *risk*, there are both upstream and downstream *mitigations* that may reduce the likelihood that malicious actors can use foundation models in harmful ways.

[2] 'Race to the bottom' entails a race to move products to market as quickly as possible, which could incentivize developer organizations to cut corners in addressing safety, security, and ethics issues. For more, see: "How Does Access Impact Risk," Institute for Security and Technology, December 2023, https://securityandtechnology.org/wp-content/uploads/2023/12/How-Does-Access-Impact-Risk-Assessing-AI-Foundation-Model-Risk-Along-A-Gradient-of-Access-Dec-2023.pdf.

[3] 'Capability overhang' occurs when a model develops capabilities and aptitudes not envisioned by an AI model's developers. Ibid.

[4] 'Malicious use' of AI models to undermine the safety and security of individuals, groups, or society includes fraud and other crime schemes, the undermining of social cohesion and democratic processes, human rights abuses, disruption of critical infrastructure, and state conflict. Ibid.

[5] 'Taking the human out of the loop' can include a model's ability to make decisions autonomously and act on them without human verification, to strategically deceive users without being instructed to do so, to self-improve or even self-replicate, and to call each others' APIs without human oversight. Ibid.

The working group collectively recognized the utility of the "upstream/downstream" framing, particularly for communicating with policymakers. Several key points about the framework were highlighted to this end:

» **Relevance to policymakers**: The framework presents a practical way to convey the implications of interventions and their potential dividends in a way that is accessible to policymakers.

» **Objective assessment**: Members appreciated the objective nature of the upstream/downstream discussion, noting that it provides a technical basis by which to develop recommendations that is less likely to be distorted by emotion or personal bias.

» **Legal considerations**: Working group members noted that, in the absence of changes to existing liability frameworks, questions could arise regarding the foreseeability of risks associated with releasing open access models. Focusing specifically on upstream risks, for example, might aid in identifying foreseeable harms within the existing legal context, thereby clarifying AI developers' duty of care in ensuring their products are safe and secure.

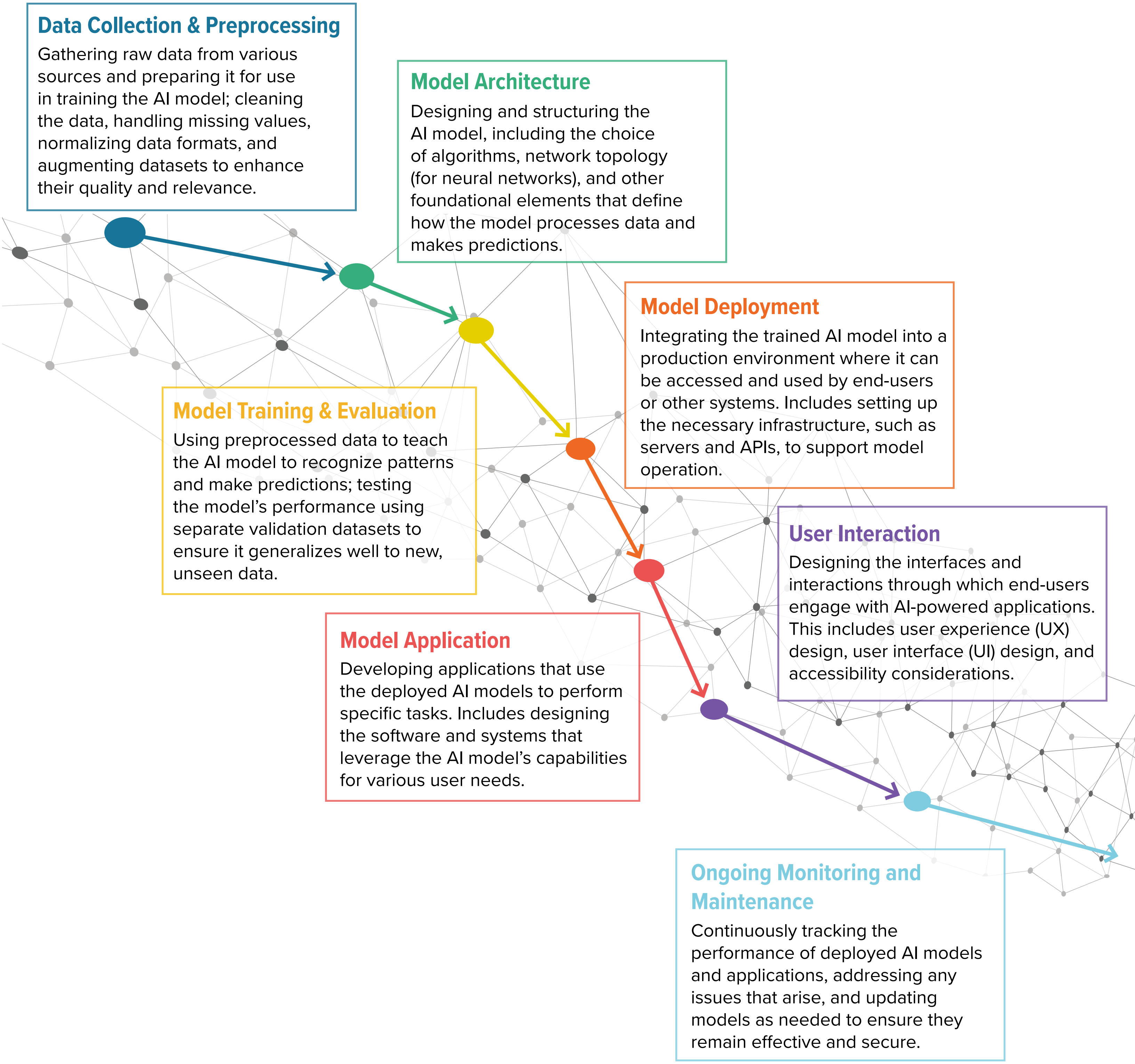
» **Risks vs. mitigations**: One member cautioned that the upstream/downstream framing requires a careful disambiguation, as both risks and their corresponding interventions could be either upstream or downstream. Working group members agreed, however, that this granularity enables the framework to provide more useful, targeted, and effective mitigation strategies.

The "upstream/downstream" framing parallels the "open vs closed" binary associated with AI proliferation in that both are necessarily reductive. Although risk mitigation in the AI sector is composed of a complex problem set, utilizing reductive framing methods to identify key areas in the ecosystem and life cycle of AI systems is useful to develop targeted risk mitigations. Building on our guiding principles, in order for a mitigation to be effective, it must be specific to, and narrowly tailored for, the relevant step in the AI lifecycle. As such, informing risk mitigation strategies through deep research and understanding of both the risks present, and the mitigation strategies available, provides policymakers and developers appropriate tools to reduce risk.

Mitigations come in two types, each equally important. As AI risks are inherently technical to some degree, given that they arise from the technology itself or user interaction with it, technical risk mitigations provide actionable and concrete changes to technical structures that may reduce risk. Policy mitigations, often in the form of regulatory frameworks and structures, provide incentives for developers to adopt successful technical mitigations at a broader scale, and can help ensure the education and safety of users.

The AI Lifecycle Stages

The AI Lifecycle Framework breaks down the complex process of AI development into manageable stages. This structured approach ensures a comprehensive understanding of each phase, making it easier to target specific risk mitigation strategies effectively.



Data Collection & Preprocessing

Gathering raw data from various sources and preparing it for use in training the AI model; cleaning the data, handling missing values, normalizing data formats, and augmenting datasets to enhance their quality and relevance.

Model Architecture

Designing and structuring the AI model, including the choice of algorithms, network topology (for neural networks), and other foundational elements that define how the model processes data and makes predictions.

Model Training & Evaluation

Using preprocessed data to teach the AI model to recognize patterns and make predictions; testing the model's performance using separate validation datasets to ensure it generalizes well to new, unseen data.

Model Deployment

Integrating the trained AI model into a production environment where it can be accessed and used by end-users or other systems. Includes setting up the necessary infrastructure, such as servers and APIs, to support model operation.

Model Application

Developing applications that use the deployed AI models to perform specific tasks. Includes designing the software and systems that leverage the AI model's capabilities for various user needs.

User Interaction

Designing the interfaces and interactions through which end-users engage with AI-powered applications. This includes user experience (UX) design, user interface (UI) design, and accessibility considerations.

Ongoing Monitoring and Maintenance

Continuously tracking the performance of deployed AI models and applications, addressing any issues that arise, and updating models as needed to ensure they remain effective and secure.

Importance of AI Lifecycle Stages

Importance of Data Collection and Preprocessing

High-quality data is crucial for training effective AI models. Poor data quality can lead to biased, inaccurate, or unreliable models, which can exacerbate risks and reduce the effectiveness of AI applications.

Importance of Model Architecture

A well-designed architecture is essential for the performance, scalability, and security of AI models. It influences how the model learns from data and its resilience to adversarial attacks or other vulnerabilities.

Importance of Model Training & Evaluation

Effective training and thorough evaluation are critical for developing accurate and reliable AI models. This stage helps identify and address issues such as overfitting, underfitting, and biases that could impact model performance in real-world applications.

Importance of Model Deployment

Deployment ensures that AI models can be utilized effectively in practical applications. Proper deployment practices are essential to maintain model performance, security, and scalability in real-world environments.

Importance of Model Application

Effective application development is crucial for ensuring that AI technologies are harnessed appropriately to provide value. This stage focuses on creating functional and efficient applications that meet user requirements and expectations.

Importance of User Interaction

Ensuring positive user interactions is essential for the successful adoption and utilization of AI technologies. This stage addresses the usability, transparency, and trustworthiness of AI applications, helping to mitigate risks associated with user misuse or misunderstanding.

Importance of Ongoing Monitoring and Maintenance

Continuous monitoring and maintenance are vital for sustaining the long-term reliability and safety of AI systems. This stage helps identify emerging risks, adapt to changing environments, and incorporate new data or advancements in AI technology.

Connection to Guiding Principles

Approaching risk mitigation through the AI Lifecycle Framework inherently aligns with our guiding principles by ensuring that interventions are feasible, actionable, and targeted. This approach necessitates the involvement of multiple stakeholders, each bringing their unique perspectives and expertise. Moreover, by targeting specific stages of the AI lifecycle, precise, informed choices can be made that protect innovation while effectively mitigating risks. Commentary on how the AI Lifestyle Framework aligns with each guiding principle follows below:

Principle #1: Balancing Innovation and Risk Aversion. The AI Lifecycle Framework allows for tailored risk mitigation strategies at each stage of AI development and deployment. This granularity ensures that risk mitigation measures are both effective and minimally intrusive, protecting the innovative potential of AI technologies. For example, focusing on upstream mitigations during the model development phase can preemptively address risks without stifling downstream innovation in application development.

Principle #2: Shared Responsibility Among Stakeholders. Effective risk mitigation requires input from all stakeholders involved in the AI lifecycle, including policymakers, developers, users, and civil society. By mapping out risks and mitigations across the entire lifecycle, each stakeholder group gains a clear understanding of their role and responsibilities. This collaborative approach leverages diverse expertise and perspectives, leading to more comprehensive and robust risk management strategies.

Principle #3: Commitment to Accuracy. Addressing risks at multiple stages of the AI lifecycle ensures that accuracy and reliability are maintained throughout the development and deployment process. Implementing rigorous validation and monitoring mechanisms at each stage upholds the integrity of AI models, thereby fostering trust and ensuring their beneficial application across various domains.

Principle #4: Feasibility-Aware Regulatory Oversight. The AI Lifecycle Framework breaks down the complex process of AI development into manageable stages, making it easier to design and implement feasible regulatory measures. By focusing on specific phases, regulators can develop targeted interventions that are both technically viable and effective, avoiding overly broad or impractical mandates.

Principle #5: Adaptable and Continuous Oversight. The dynamic nature of the AI Lifecycle Framework aligns with the need for adaptable regulatory frameworks that evolve with technological advancements. Continuous oversight and iterative improvements are built into the lifecycle, allowing for regular updates to risk mitigation strategies based on emerging threats and new developments in AI technology.

The Framework in Conclusion

The AI Lifecycle Framework provides a comprehensive and systematic approach to identifying and mitigating risks associated with AI foundation models. By breaking down the complex process of AI development into manageable stages, this framework enables targeted interventions that are both effective and minimally intrusive. The approach aligns closely with the guiding principles, ensuring that risk mitigation strategies are feasible, actionable, and balance innovation with risk aversion.

A key strength of the AI Lifecycle Framework is its emphasis on deep research and informed decision-making. By thoroughly examining each stage of the AI lifecycle, it is possible to develop a nuanced understanding of the specific risks and opportunities associated with AI foundation models. This research-driven approach allows for the identification of targeted mitigations that are grounded in technical expertise and real-world evidence, rather than relying on broad generalizations or untested assumptions.

Another important aspect of the framework is its recognition of the spectrum of openness in AI development and deployment. Open access AI models, more freely available for use, modification, and distribution, present both unique opportunities and risk mitigation challenges. While the open access nature of these models promotes transparency, collaboration, and rapid innovation, it also means that they can be more easily accessed and potentially misused by malicious actors.

The AI Lifecycle Framework takes these considerations into account, acknowledging that certain risk mitigations may be more effective for models across the openness spectrum, while others may be more applicable to closed-source systems where there is greater control over a model’s access and use. For example, imposing restrictions on the downstream use of models further down the openness gradient may be less effective, as the open and distributed nature of these models makes centralized enforcement more challenging. Instead, the AI Lifecycle Framework suggests focusing on upstream mitigations for open access models, such as ensuring responsible data collection and preprocessing, requiring pre-deployment red teaming, and promoting transparency and accountability in model development.

By considering the distinct challenges and opportunities associated with open access models, the AI Lifecycle Framework enables policymakers and developers to craft targeted, context-specific risk mitigation strategies. Instead of developing separate risk mitigation strategies for open vs. closed models, the AI Lifecycle Framework focuses on developing effective mitigation targeting for models located at any point across the openness spectrum. This nuanced approach ensures that the benefits of open access AI development are maximized while potential risks are effectively managed..

Deep Dive on Malicious Use Risks

The concept of “Malicious Use” in AI is a challenging one to tackle, as there are many avenues through which malicious actors could utilize AI technology. For the purposes of this research, we focus on several key areas identified in our Phase I report.

TYPES OF MALICIOUS USE RISKS

Fraud and other crime schemes enabled by AI-generated social engineering content, particularly when targeting at-risk populations (e.g., children and the elderly).

The undermining of social cohesion and democratic processes through targeted disinformation campaigns that seek to sow discord or confusion, particularly within the context of elections and political transitions.

Human rights abuses by expanding the ability of authoritarian states to surveil, constrain, and oppress minorities and dissidents.

Disruption of critical infrastructure by providing malicious actors with offensive cyber capabilities that outmatch defenses or by introducing cybersecurity vulnerabilities to critical systems.

State conflict by contributing to the capabilities of adversarial or revanchist states looking for the means to overcome power and information asymmetries, including through economic, military, intelligence, cyber, cognitive, and/or information operations.

METHODOLOGY

For each of our five malicious use cases, we first examine overarching trends of these malicious behaviors without the use of AI technology to build an understanding of the ecosystems and behavioral patterns that surround each category. Second, we examine the current status of malicious use of AI technologies in each category to determine how these new AI technologies are being applied to existing behavioral patterns. Third and finally, we take a future facing approach, to determine how, with extant and more advanced technologies, AI tools might be applied to each category.

This report seeks to establish a clear historical context for each malicious use case so that readers may introduce AI systems to the mix. The objective here is twofold. First, to understand how AI systems are currently fitting into existing patterns of human behavior in malicious use cases. Second, to build a predictive mental model to determine where and how AI systems may fit into, or exacerbate, existing patterns of malicious activity, and to identify the most likely and most threatening potential outcomes.

Understanding both the historical and current patterns of malicious behaviors enables us to make predictions about the future. While these potential scenarios underscore the need for proactive governance and the development of robust AI safety frameworks, it is crucial to approach such predictions with caution. The rapid pace of technological advancement and the complex interplay of human choices and actions introduce significant uncertainty into any long-term forecasts.

Fraud and Other Crime Schemes

Our Phase I report identified “fraud and other crime schemes” as a subcategory of the risk of malicious use, which can be enabled by AI-generated social engineering content, particularly when targeting at-risk populations (e.g., children and the elderly).

Historical Perspective

Fraud and criminal activities have long plagued human societies, with perpetrators employing various tactics to exploit vulnerabilities and deceive their victims. Conventional fraud tactics, such as identity theft, Ponzi schemes, and confidence tricks, have been used for centuries to manipulate individuals and organizations for financial gain.[6] These methods often rely on social engineering techniques, which involve psychological manipulation to trick people into divulging sensitive information or making financial transactions.[7]

Mass marketing fraud, which emerged in the 20th century, saw the use of telemarketing, mail fraud, and early forms of email scams to reach wider audiences.[8] These tactics laid the

[6] Michael Levi, “Organized Fraud and Organizing Frauds: Unpacking research on networks and organization,” *Criminology & Criminal Justice*, 8 no. 4 (2008): 389-419, https://doi.org/10.1177/1748895808096470.

[7] Michael Workman, “Gaining Access with Social Engineering: An Empirical Study of the Threat,” *Information Systems Security*, 16 no. 6 (2007): 315–31, https://doi.org/10.1080/10658980701788165.

[8] Mark Button, Chris Lewis, and Jacki Tapley, “Fraud Typologies and the Victims of Fraud: Literature Review,” University of Portsmouth, 2009, https://researchportal.port.ac.uk/en/publications/fraud-typologies-and-the-victims-of-fraud-literature-review.

groundwork for more sophisticated digital fraud schemes that would later emerge with the advent of AI technologies. By studying the historical evolution of fraud and crime schemes, we can better understand the underlying human behaviors and motivations that drive these activities and develop proactive measures to counter them.

Current State of Play

Fraud and other crime schemes are increasingly leveraging AI technologies. AI-generated phishing emails and messages have surged, with a 1,265 percent increase in malicious phishing emails and 967 percent rise in credential phishing in particular since the fourth quarter of 2022, according to a report by cybersecurity firm SlashNext.[9] Deepfake technology is being exploited for identity theft and fraud. In one case, a finance worker at a multinational firm was tricked into paying out $25 million to fraudsters using deepfake technology to pose as the company's CFO in a video call.[10] In 2019, a UK company employee was conned into sending $240,000 to cybercriminals who used AI voice cloning to impersonate the head of the company in a phone call.[11] The Asia-Pacific region saw a 1530 percent increase in deepfake fraud cases between 2022 and 2023, the second highest globally, with Vietnam, Japan, and the Philippines experiencing some of the biggest increases.[12] Voice cloning has even been used for "false ransoms" claiming kidnapped children, as seen in a disturbing Arizona case where a mother received a convincing ransom call from her daughter, who was, in reality, safe and sound.[13] Multiple examples of custom designed AI systems for fraudulent behavior have been identified in underground Russian forums, demonstrating a growing business model surrounding fraudulent use of AI technologies.[14]

Outlook

As AI-generated content continues to advance in sophistication, criminals may exploit these capabilities to devise highly targeted and persuasive scams aimed at vulnerable populations. By leveraging AI to analyze vast amounts of personal data, including social media activity, online behavior patterns, and individual circumstances, perpetrators could craft meticulously personalized phishing emails or deepfake images, audio, or video designed to exploit the

[9] "The State of Phishing 2023," SlashNext, April 10, 2024, https://slashnext.com/state-of-phishing-2023/.

[10] Heather Chen and Kathleen Magramo, "Finance worker pays out $25 million after video call with deepfake 'chief financial officer,'" *CNN*, February 4, 2024, https://www.cnn.com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intl-hnk/index.html.

[11] Catherine Stupp, "Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case," *The Wall Street Journal*, August 30, 2019, https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402.

[12] Natnicha Surasit, "Rogue Replicants: Criminal Exploitation of Deepfakes in South East Asia," Global Initiative Against Transnational Organized Crime, February 29, 2024, https://globalinitiative.net/analysis/deepfakes-ai-cyber-scam-south-east-asia-organized-crime/.

[13] Ben Cost, "AI Clones Teen Girl's Voice in $1M Kidnapping Scam: 'I've Got Your Daughter'," *New York Post*, April 12, 2023, https://nypost.com/2023/04/12/ai-clones-teen-girls-voice-in-1m-kidnapping-scam/.

[14] Check Point Team, "Generative AI Is the Pride of Cybercrime Services," Check Point, February 1, 2024, https://blog.checkpoint.com/research/generative-ai-is-the-pride-of-cybercrime-services/.

specific vulnerabilities of each target. Moreover, AI technology may enable the orchestration of large-scale, complex fraud schemes by automating key aspects of the criminal operation, from identifying potential victims to obscuring the illicit proceeds. Furthermore, additional openness poses additional risk in these sectors, as the use of open access models would invalidate the capabilities of model builders to rescind access to models when they are utilized for fraud or other crime schemes.

The Undermining of Social Cohesion and Democratic Processes

Our Phase I report identified "the undermining of social cohesion and democratic processes" as a subcategory of the risk of malicious use. AI technologies, through targeted disinformation campaigns, may seek to sow discord or confusion, particularly within the context of elections and political transitions.

Historical Perspective

The use of propaganda and manipulation of media to influence public opinion and undermine democratic processes has a long and complex history. Throughout the 20th century, authoritarian regimes and political groups employed various techniques to control narratives and suppress dissent. These methods included the ownership and control of media outlets, censorship of opposing viewpoints, and the creation of state-sponsored propaganda to shape public perceptions.[15] In addition to media manipulation, historical instances of election interference, such as ballot stuffing, voter suppression, and the spread of misinformation through word of mouth or printed materials, demonstrate the enduring nature of attempts to subvert democratic processes.[16] By examining these historical tactics, we can gain a deeper understanding of the human behaviors and power dynamics that contribute to the erosion of social cohesion and the weakening of democratic institutions. This knowledge is crucial in developing effective strategies to counter the potential misuse of AI technologies in the realm of information warfare and political manipulation.

[15] Garth S. Jowett et al., *Propaganda and Persuasion*, 5th ed. (SAGE, 2012), https://hiddenhistorycenter.org/wp-content/uploads/2016/10/PropagandaPersuasion2012.pdf.

[16] Fabrice Lehoucq, "Electoral Fraud: Causes, Types, and Consequences," *Annual Review of Political Science* 6, no. 1 (June 1, 2003): 233–56, https://doi.org/10.1146/annurev.polisci.6.121901.085655.

Current State of Play

Social media bots and inauthentic automated accounts are being employed to spread misinformation. While early bot manipulation of social networks did not use AI, experts have uncovered evidence demonstrating AI systems are now contributing to existing bot networks. OpenAI reported that threat actors based in Russia, China, Iran, and Israel utilized their models "to generate social media comments in multiple languages, make up names and bios for fake accounts, create cartoons and other images, and debug code," according to May 2024 NPR reporting.[17] Although these actors were utilizing OpenAI's models, OpenAI noted that the actors struggled to receive natural engagement on AI generated content. Regardless of this failure, clear evidence now demonstrates that malicious use of AI systems to undermine social cohesion and democratic processes is underway. Additionally, AI-powered sentiment analysis is being leveraged to manipulate public opinion. A June 2023 report found that market success "depend[s] on the effectiveness of AI implementation and continuous improvement of the quality of opinion mining and semantic recognition and analysis."[18] While the report focuses on AI in marketing, it notes the key factor is manipulating public opinion, which bad actors can exploit to influence elections and reduce social cohesion. Deepfakes and synthetic media are being used for political manipulation. In January 2024, CNN reported on a robocall using an AI voice resembling President Biden that advised New Hampshire residents against voting in the presidential primary.[19] In December 2023, AP News reported that AI-generated audio recordings impersonated a Slovak liberal candidate discussing plans to raise beer prices and rig the election, which spread on social media despite fact-checkers identifying them as false.[20] A Stanford/University of Chicago whitepaper highlighted how politicians may dismiss genuine information by undermining the credibility of the information environment,[21] citing a compromising video of Turkish candidate Muharrem İnce, who claimed it was a deepfake before withdrawing.[22]

[17] Shannon Bond, "In a first, OpenAI removes influence operations tied to Russia, China, and Israel," *NPR*, May 31, 2024, https://www.npr.org/2024/05/30/g-s1-1670/openai-influence-operations-china-russia-israel.

[18] Michael Gerlich, Walaa Elsayed, and Konstantin Sokolovskiy, "Artificial Intelligence as Toolset for Analysis of Public Opinion and Social Interaction in Marketing: Identification of Micro and Nano Influencers," *Frontiers in Communication* no. 8 (June 15, 2023), https://doi.org/10.3389/fcomm.2023.1075654.

[19] Em Steck and Andrew Kacinzsky, "Fake Joe Biden robocall urges New Hampshire voters not to vote in Tuesday's Democratic primary,'" *CNN*, January 22, 2024, https://edition.cnn.com/2024/01/22/politics/fake-joe-biden-robocall/index.html.

[20] Ali Swenson and Christine Fernando, "Misinformation May Get Worse in 2024 Election as Safeguards Erode," *AP News*, December 26, 2023, https://apnews.com/article/election-2024-misinformation-ai-social-media-trump-6119ee6f498db10603b3664e9ad3e87e.

[21] Ethan Bueno De Mesquita et al., "Preparing for Generative AI in the 2024 Election: Recommendations and Best Practices Based on Academic Research," white paper, University of Chicago Harris School of Public Policy and the Stanford Graduate School of Business, 2023, https://harris.uchicago.edu/files/ai_and_elections_best_practices_no_embargo.pdf.

[22] Nicolas Camut, "Fresh Blow for Erdoğan, as Rival Pulls Out of Turkey Election Amid Sex Tape Scandal," *POLITICO*, May 11, 2023, https://www.politico.eu/article/fresh-blow-for-erdogan-as-rival-pulls-out-of-turkey-election-amid-sex-tape-scandal/.

Outlook

The continued development of advanced AI language models could present a growing threat to the integrity of democratic processes and social cohesion. While OpenAI effectively identified and halted malicious use of their models by state actors, the makers of open access models do not share the capability to do so. As developers increase access to their models, they become more susceptible to fine tuning and repurposing than closed access models, and have the capability to be harnessed by malicious actors to generate and disseminate vast amounts of targeted disinformation and propaganda across multiple platforms simultaneously, making it increasingly difficult for the public to discern truth from falsehood. Furthermore, the potential misuse of AI to create hyper-realistic deepfakes depicting politicians and public figures in fabricated scenarios could have far-reaching implications for the democratic process. The strategic deployment of such deepfakes, particularly during sensitive periods such as elections, could significantly influence voter opinions and potentially manipulate electoral outcomes.

Human Rights Abuses

Our Phase I report identified "human rights abuses" as a subcategory of the risk of malicious use. AI technologies are expanding the ability of authoritarian states to surveil, constrain, and oppress minorities and dissidents.

Historical Perspective

Throughout history, authoritarian regimes and oppressive governments have employed various tactics to control and suppress their populations. Prior to the widespread adoption of the internet and other connected technologies, surveillance states relied on a network of informants, postal interception, and manual surveillance techniques to monitor and gather information on citizens.[23] These methods, while labor-intensive and often limited in scope, served as powerful tools for maintaining control and quelling dissent. Political repression, another common tactic, involved the suppression of opposition voices, the imprisonment of political opponents, and the use of torture to extract confessions or information.[24] These actions, often carried out by secret police or military forces, created a climate of fear and compliance, enabling regimes to maintain their grip on power.

[23] Anna Funder, *Stasiland: Stories from Behind the Berlin Wall* (Granta Books, 2003), https://granta.com/products/stasiland/?binding=ebook.

[24] Robert Justin G. Goldstein, *Political Repression in Modern America: From 1870 to 1976* (University of Illinois Press, 2001).

The study of historical cases of political repression provides valuable insights into the psychological and societal impacts of these tactics, informing our understanding of how AI-powered surveillance and control mechanisms might be misused by malicious actors. Ethnic and political cleansing—the systematic targeting of specific groups based on their race, religion, or political views—represents one of the most egregious forms of human rights abuse. Historical examples, such as the Holocaust, the Rwandan genocide, and the persecution of Uyghurs in China, demonstrate the devastating consequences of state-sponsored violence and discrimination. By examining the ideologies, propaganda, and logistical mechanisms that enabled these atrocities, we can better understand the potential risks posed by AI technologies in the hands of malicious actors seeking to target and harm specific populations.

Current State of Play

In China, an extensive network of facial recognition cameras and AI-powered surveillance systems track citizens, particularly targeting the Uyghur Muslim minority, enabling abuses like mass detention, forced labor, and movement restrictions.[25] The Electronic Frontier Foundation's Atlas of Surveillance project tracks the use of surveillance tools throughout the United States— including facial recognition and AI-enabled policing activities—where disparate procurement and use with limited oversight might threaten human rights.[26] Predictive policing algorithms and risk assessment tools from companies like Palantir and Geolitica have faced fierce backlash amid accusations of inaccuracy, racism, and bias.[27,28,29,30] An October 2023 analysis from The Markup found Geolitica's flagship predictive policing technology was accurate in less than one half of one percent of cases (n=23,631).[31] AI is assisting censorship and content moderation; experts believe China is already using AI to further develop oversight mechanisms for internet traffic and use.[32]

[25] Paul Mozur, "One Month, 500,000 Face Scans: How China Is Using A.I. to Profile a Minority," *The New York Times*, April 14, 2019, https://www.nytimes.com/2019/04/14/technology/china-surveillance-artificial-intelligence-racial-profiling.html.

[26] "Atlas of Surveillance," Electronic Frontier Foundation and the University of Nevada, Reno Reynolds School of Journalism, last updated March 8, 2024, https://atlasofsurveillance.org/atlas.

[27] Ali Winston, "Palantir Has Secretly Been Using New Orleans to Test Its Predictive Policing Technology," *The Verge*, February 27, 2018, https://www.theverge.com/2018/2/27/17054740/palantir-predictive-policing-tool-new-orleans-nopd.

[28] Art Raymond, "High-tech Surveillance Company Banjo Was Never Able to Do What It Claimed, Audit Finds," *Deseret News*, December 20, 2023, https://www.deseret.com/utah/2021/4/14/22375665/utah-banjo-surveillance-personal-privacy-white-supremacist-audit-law-enforcement-kkk/.

[29] Will Douglas Heaven, "Predictive Policing Algorithms Are Racist. They Need to Be Dismantled," *MIT Technology Review*, June 21, 2023, https://www.technologyreview.com/2020/07/17/1005396/predictive-policing-algorithms-racist-dismantled-machine-learning-bias-criminal-justice/.

[30] SoundThinking, "Resource Management," October 11, 2023, https://www.soundthinking.com/law-enforcement/resource-deployment-resourcerouter/.

[31] Aaron Sankin and Surya Mattu, "Predictive Policing Software Terrible At Predicting Crimes," *The Markup*, October 2, 2023, https://themarkup.org/prediction-bias/2023/10/02/predictive-policing-software-terrible-at-predicting-crimes.

[32] Sarah Zheng, "China's Answers to ChatGPT Have a Censorship Problem," *Bloomberg*, May 2, 2023, https://www.bloomberg.com/news/newsletters/2023-05-02/china-s-chatgpt-answers-raise-questions-about-censoring-generative-ai.

Outlook

The integration of AI technology into the surveillance apparatus of authoritarian regimes poses a significant threat to human rights. By harnessing the power of AI, regimes could construct even more pervasive and oppressive surveillance states capable of continuously monitoring citizens' digital footprints, analyzing their behaviors and associations, and proactively identifying potential dissidents for targeting. Additionally, AI could be employed to automate censorship on an unprecedented scale, instantaneously identifying and suppressing any online content that challenges the regime's authority or narratives. Over time, the persistent application of such AI-driven censorship could fundamentally reshape citizens' perceptions of reality and further entrench the power of oppressive regimes.

Disruption of Critical Infrastructure

Our Phase I report identified "disruption of critical infrastructure" as a subcategory of the risk of malicious use. AI can provide malicious actors with offensive cyber capabilities that outmatch defenses or can introduce cybersecurity vulnerabilities to critical systems.

Historical Perspective

The disruption of critical infrastructure, such as energy grids, transportation networks, and communication systems, has long been a tactic employed by both state and non-state actors to weaken adversaries and sow chaos. Historical examples of sabotage, particularly during wartime, demonstrate the vulnerability of civilian critical infrastructure to physical attacks. The bombing of railways, bridges, and industrial facilities, for instance, was a common tactic used to cripple an enemy's ability to wage war and maintain economic stability.[33] Economic blockades and sanctions, another form of infrastructure disruption, have been used by nations to exert pressure on adversaries by cutting off access to essential goods and services. These tactics, while not physically destructive, can have severe consequences for the targeted country's economy and population.

The examination of historical cases of economic warfare provides valuable insights into the potential risks posed by AI-powered attacks on critical infrastructure, particularly in the realm of cyber warfare. Technological espionage, the theft or sabotage of proprietary technologies and intellectual property, has a long history in the context of state conflict. The Cold War, for example, saw extensive efforts by both the United States and the Soviet Union to acquire

[33] Scott Gerwehr and Russell W. Glenn, *The Art of Darkness: Deception and Urban Operations* (RAND, 2000), https://www.rand.org/content/dam/rand/pubs/monograph_reports/MR1132/RAND_MR1132.pdf.

each other's technological secrets, often through human intelligence and covert operations.[34] The study of these historical cases highlights the enduring importance of protecting critical technologies and the potential for AI to be used as a tool for industrial espionage and sabotage.

Current State of Play

AI-powered malware and cyber-attacks have become increasingly sophisticated, with threat actors integrating proven strategies with AI tools to enhance their capabilities. Cyber attacks on critical infrastructure are no new phenomenon, with the Stuxnet attack and Energetic Bear demonstrating the capability of cyber attacks to affect core infrastructure.[35,36] Cyber attacks are increasingly targeting water systems in the United States; according to a recent White House warning to Governors, threat actors affiliated with the Iranian government targeted and disabled operational technology used at water facilities in the United States.[37] The letter also warned that Chinese government sponsored "Volt Typhoon" actors are pre-positioning themselves in advance to disrupt critical infrastructure operations in the United States if geopolitical tensions or military conflicts arise.[38] A Russian military intelligence group variously nicknamed Forest Blizzard/APT28/Fancy Bear/Strontium, known to support Russia's foreign policy and military goals, utilized AI tools for research into satellite and radar technologies relevant to military operations in Ukraine and for general cyber operation support.[39]

Use of AI by bad actors to evade security systems and defenses has emerged as a potent threat; IBM researchers created DeepLocker, a form of AI-powered malware that uses machine learning to target specific victims and avoid detection by traditional security measures.[40] DeepLocker only activates when it recognizes certain criteria, making it much harder to defend against. Similar to the fraud examples described above, AI-driven social engineering can be used in attacks on infrastructure, as AI enables highly tailored spear-phishing by

[34] Kristie Macrakis, "Seduced by Secrets: Inside the Stasi's Spy-Tech World," *The American Historical Review* 114, no. 4 (October 1, 2009): 1181–82, https://doi.org/10.1086/ahr.114.4.1181.

[35] Ellen Nakashima and Joby Warrick, "Stuxnet Was Work of U.S. and Israeli Experts, Officials Say," *The Washington Post*, May 20, 2023, https://www.washingtonpost.com/world/national-security/stuxnet-was-work-of-us-and-israeli-experts-officials-say/2012/06/01/gJQAlnEy6U_story.html.

[36] Nicole Perlroth, "Russian Hackers Targeting Oil and Gas Companies", *The New York Times*, June 30, 2014, https://www.nytimes.com/2014/07/01/technology/energy-sector-faces-attacks-from-hackers-in-russia.html?partner=slack&smid=sl-share.

[37] Michael S. Regan and Jake Sullivan, "Letter to Governor Regarding Cyberattacks on Water Systems," Environmental Protection Agency and the White House, March 18, 2024, https://www.epa.gov/system/files/documents/2024-03/epa-apnsa-letter-to-governors_03182024.pdf.

[38] "Chinese Government Poses 'Broad and Unrelenting' Threat to U.S. Critical Infrastructure, FBI Director Says," Federal Bureau of Investigation, April 18, 2024, https://www.fbi.gov/news/stories/chinese-government-poses-broad-and-unrelenting-threat-to-u-s-critical-infrastructure-fbi-director-says.

[39] Microsoft Threat Intelligence, "Staying Ahead of Threat Actors in the Age of AI," Microsoft Security Blog, February 27, 2024, https://www.microsoft.com/en-us/security/blog/2024/02/14/staying-ahead-of-threat-actors-in-the-age-of-ai/.

[40] Dhilung Kirat et al., "DeepLocker," IBM Research, 2003, https://i.blackhat.com/us-18/Thu-August-9/us-18-Kirat-DeepLocker-Concealing-Targeted-Attacks-with-AI-Locksmithing.pdf.

analyzing a target's digital footprint to craft persuasive messages.[41] In 2023, AI-powered phishing increased by over 1200 percent compared to the previous year. A North Korean threat actor, variously nicknamed Emerald Sleet/Thallium/Kimsuky/Velvet Chollima, known for spear-phishing prominent individuals with expertise on North Korea, leveraged AI tools to understand public vulnerabilities and drafting content for spear-phishing campaigns.[42] By studying public data, AI can identify key personnel with access to critical systems and launch attacks tailored to these high-value targets.

Outlook

Sophisticated AI systems could be designed to infiltrate and learn the intricacies and vulnerabilities of essential networks, such as power grids, and subsequently coordinate precise, devastating attacks that result in widespread disruption and damage. Moreover, AI could be utilized to automate the target selection process for infrastructure attacks by analyzing vast datasets to identify the most susceptible and high-impact targets, enabling attackers to optimize the destructive impact of their campaigns.

State Conflict

Our Phase I report identified "state conflict" as a subcategory of the risk of malicious use. AI can contribute to the capabilities of adversarial or revanchist states looking for the means to overcome power and information asymmetries, including through economic, military, intelligence, cyber, cognitive, and/or information operations.

Historical Perspective

The history of state conflict is marked by a wide range of tactics and strategies employed to gain military, economic, and political advantages over adversaries. The study of conventional warfare, including the use of espionage, propaganda, and economic sanctions, provides a foundation for understanding how AI technologies might be leveraged in future conflicts. The use of psychological warfare, the manipulation of information to demoralize and confuse enemy forces and populations, has been a staple of state conflict for centuries. Historical examples, such as the use of leaflet drops and radio broadcasts during World War II, demonstrate the power of targeted information campaigns to influence the course of conflicts.[43]

[41] HRSS CPAs, "Cyber Defense Adversarial AI for Government Security," HRSS CPAs (blog), July 20, 2023, https://hrss.cpa/adversarial-ai-cyber-defense-government-security-digital/.

[42] Microsoft Threat Intelligence, "Staying Ahead of Threat Actors."

[43] Paul M. A. Linebarger, *Psychological Warfare*, 2nd ed. (Duell, Sloan and Pearce, 1969), https://www.gutenberg.org/files/48612/48612-h/48612-h.htm.

The examination of these tactics informs our understanding of how AI-powered information warfare might be used to shape public opinion and undermine adversaries in future conflicts. As the boundaries between physical and digital conflict continue to blur, the study of historical cases of technological espionage and sabotage becomes increasingly relevant. The Stuxnet cyberattack on Iran's nuclear program, for example, demonstrated the potential for digital weapons to cause physical damage and disruption.[44] By examining these historical cases, we can better anticipate the potential risks posed by AI-powered cyberattacks and develop effective strategies for defending against them.

Current State of Play

AI is already being utilized in state conflict. In 2020, a United Nations report documented the use of the Turkish-manufactured STM Kargu-2 autonomous drone in Libya to hunt down and attack retreating enemy forces without human intervention, believed to be the first such case.[45] According to a 2024 DefenseOne report, the Pentagon is testing AI-enabled weapons from aerial drones to autonomous seafaring vehicles.[46] The Ukraine conflict has spurred development of jam-resistant drones that use AI to operate when jammed.[47] Ukrainian drone company Saker has launched AI systems for autonomous strikes and intelligence gathering against Russian forces. And U.S. defense company Anduril is building next generation information gathering and analysis tools to autonomously "detect, track, and classify every object of interest in an operator's vicinity" and deliver warfighters an intelligent common operating picture.[48]

As with criminal actors, states will utilize AI cyberwarfare tools to conduct military missions and identify weaknesses in opposing systems, on and off the battlefield. An Iranian cyber threat actor linked to the Islamic Revolutionary Guard Corps (IRGC), variously nicknamed Crimson Sandstorm/Curium/Tortoiseshell/Imperial Kitten/Yellow Liderc, has been observed targeting defense, maritime shipping, transportation, healthcare, and technology sectors using watering hole attacks and social engineering to deploy malware.[49] The actors leveraged AI tools to create phishing emails and lure targets to malicious websites, generate code for app and web development, remote server interactions, web scraping, and system information extraction.

[44] Kim Zetter, *Countdown to Zero Day - Stuxnet and the Launch of the World's First Digital Weapon* (Crown, 2014), https://www.academia.edu/43420700/Countdown_to_Zero_Day_Stuxnet_and_the_Launch_of_the_Worlds_First_Digital_Weapon.

[45] "Final report of the Panel of Experts on Libya established pursuant to Security Council resolution 1973 (2011)," United Nations Security Council, March 8, 2021, https://undocs.org/Home/Mobile?FinalSymbol=S%2F2021%2F229&Language=E&DeviceType=Desktop&LangRequested=False.

[46] Patrick Tucker, "The Pentagon Is Already Testing Tomorrow's AI-powered Swarm Drones, Ships," *Defense One*, January 23, 2024, https://www.defenseone.com/technology/2024/01/pentagon-already-testing-tomorrows-ai-powered-swarm-drones-ships/393528/.

[47] David Hambling, "Ukraine's AI Drones Seek and Attack Russian Forces Without Human Oversight," *Forbes*, October 17, 2023, https://www.forbes.com/sites/davidhambling/2023/10/17/ukraines-ai-drones-seek-and-attack-russian-forces-without-human-oversight/?sh=71f4445a66da.

[48] "Anduril," Command & Control, accessed May 2024, https://www.anduril.com/command-and-control/.

[49] Microsoft Threat Intelligence, "Staying Ahead of Threat Actors."

They also used large language models to develop techniques to evade detection and disable antivirus.

An official from the South Korean National Intelligence Service reported that North Korean threat actors are actively using generative AI tools to conduct sophisticated cyberattacks and identify hacking targets, further integrating AI into state conflict.[50]

Outlook

The application of AI in the development of advanced weapons systems raises alarming prospects for the future of state conflict. The deployment of highly autonomous, AI-powered weapons, such as drone swarms capable of overwhelming defenses and executing coordinated strikes without human intervention, could significantly lower the threshold for armed confrontation between nations. In addition, AI could be weaponized to wage unprecedented information warfare campaigns against adversaries. By generating personalized propaganda tailored to each individual's digital profile, with content algorithmically optimized to erode trust in institutions and amplify societal divisions, malicious actors could unleash large-scale disinformation campaigns designed to destabilize entire societies. As openness increases, the barrier to entry for smaller state and non-state actors seeking to develop or utilize advanced AI capabilities for military applications is lowered. For state actors without the capability to internally develop advanced AI models, higher degrees of openness provides the capability to initialize AI research from a running start instead of starting from a standstill.

Applying the Framework

In this section, we leverage our guiding principles to apply our AI Lifecycle Framework to the identified risk category: Malicious Use. We systematically examine each phase of the AI lifecycle and propose technical or policy-based risk mitigation strategies. Although the following mitigation approaches are tailored to address this single risk category, they may also contribute to broader risk mitigation efforts, thereby reducing other associated risks. It is important to recognize that the effectiveness of risk mitigation strategies may vary across different stages of the AI lifecycle. In some stages, certain mitigations may prove more effective, while in others, there may be no viable strategies to address the identified risk.

AI models released in more open manners present a distinct set of risks and considerations when it comes to mitigating malicious use. The open and accessible nature of these models

[50] Jayant Chakravarti, "North Korean Hackers Using AI in Advanced Cyberattacks," *Data Breach Today*, January 24, 2024, https://www.databreachtoday.com/north-korean-hackers-using-ai-in-advanced-cyberattacks-a-24184.

means that they can be more easily obtained, modified, and repurposed by malicious actors, making it more challenging to control their use and prevent misuse. At the same time, the open access approach promotes transparency, collaboration, and rapid innovation, which can be harnessed to develop more robust and adaptable risk mitigation strategies.

It is important to recognize that the effectiveness of risk mitigation strategies may vary across different stages of the AI lifecycle, and that in some cases, the open access nature of a model may limit the feasibility of certain mitigations.

In instances where effective risk mitigation strategies are lacking, we advise against the inclusion of superficial or redundant measures. Such an approach would undermine key principles, including the balance between innovation and risk aversion, feasibility-aware regulatory structures, and adaptable regulatory frameworks. The inclusion of ineffective mitigations not only detracts from the integrity of our recommendations but also conflicts with the foundational principles guiding our risk management strategies. Therefore, our focus remains on proposing robust, actionable, and contextually appropriate mitigations that align with our guiding principles.

The figures contained in the following five pages depict the working group's recommended mitigations associated with the relevant stage of the AI lifecycle and tagged with the risk or risks that it may be effective in mitigating.

	Data Collection and Preprocessing	Types of risks mitigated
policy	**Dataset sourcing transparency** Require dataset sourcing transparency for large labs building foundation models, so that users and civil society can see where datasets originated. ➔ *Transparency in data sourcing helps prevent the use of data from unethical sources, which can be used for manipulative or oppressive purposes.*	Human rights abuses Undermining social cohesion and democratic processes
technical	**Data validation and sanitization** Implement rigorous data validation and sanitization protocols to detect and remove anomalous or suspicious data points before they enter the training pipeline. ➔ *Sanitization protocols ensure that malicious data is removed early, preventing it from affecting the model's behavior or integrity.*	Fraud and crime schemes Disruption of critical infrastructure
technical	**Privacy-preserving AI techniques** Employ privacy-preserving AI techniques, such as federated learning and secure multi-party computation, to protect sensitive data and prevent malicious exploitation. ➔ *These techniques protect sensitive data from being exposed or misused, reducing the risk of data being leveraged for malicious purposes.*	Human rights abuses Fraud and crime schemes

	Model Architecture	Types of risks mitigated
policy	**AI roundtables** Develop and support AI roundtables for verified researchers from leading labs, top universities, and American startups to diffuse best practices and methods throughout the American AI ecosystem. ➔ *Knowledge sharing promotes secure development practices, reducing vulnerabilities that can be exploited in conflicts or critical infrastructure attacks.*	State conflict Disruption of critical infrastructure
policy	**Robust security standards** Develop and require robust security standards for leading labs to preserve operational security to prevent theft of IP and leading models in closed systems. ➔ *Security standards help protect intellectual property and models from being stolen and misused in state conflicts or attacks on infrastructure.*	State conflict Disruption of critical infrastructure

Key: Types of Risks Mitigated

- ⚠ Human rights abuses
- Fraud and crime schemes
- Disruption of critical infrastructure
- ? Undermining social cohesion and democratic processes
- State conflict
- AI risk as a whole

	Model Architecture (cont.)	Types of risks mitigated
policy	**Cash, compute, or grant incentives** Provide cash, compute or grant incentives to organizations and researchers who participate in collaborative projects aimed at sharing knowledge and best practices in secure AI development. → *Collaboration fosters the development of secure AI technologies, mitigating risks associated with misuse in conflicts and oppressive regimes.*	🌍 ⚠
policy	**Legal protections and reward programs for whistleblowers** Implement strong legal protections and reward programs for whistleblowers who report unethical or malicious use of AI technologies within organizations. → *Protecting whistleblowers encourages the reporting of unethical or unsafe practices, helping to prevent malicious uses of AI.*	☼
technical	**Privacy-preserving AI techniques** Employ privacy-preserving AI techniques, such as federated learning and secure multi-party computation, to protect sensitive data and prevent malicious exploitation. → *Protecting sensitive data reduces the risk of it being exploited for malicious purposes, such as fraud or surveillance.*	⚠ ⛓

	Model Training and Evaluation	Types of risks mitigated
policy	**Regular security audits and penetration testing** Mandate regular security audits and penetration testing of AI training environments to identify and address vulnerabilities that could be exploited for malicious purposes. → *Regular security checks help identify and fix vulnerabilities, preventing exploitation during lifecycle stages after training and deployment.*	⛓ 🏭
policy	**Strong legal protections and reward programs for whistleblowers** Implement strong legal protections and reward programs for whistleblowers who report unethical or malicious use of AI technologies within organizations. → *Protecting whistleblowers encourages the reporting of unethical or unsafe practices, helping to prevent malicious uses of AI.*	☼

Key: Types of Risks Mitigated

⚠	Human rights abuses
⛓	Fraud and crime schemes
🏭	Disruption of critical infrastructure
🗳	Undermining social cohesion and democratic processes
🌍	State conflict
☼	AI risk as a whole

	Model Training and Evaluation (cont.)	Types of risks mitigated
policy	Bug bounty programs Create bug bounty programs to identify weaknesses in known methodologies. ➜ *As bug bounties open up new financial motivations for finding and solving bugs, new approaches and risk mitigations may arise.*	☼
technical	Red teaming Conduct red team, including, but not limited to, adversarial testing to simulate and identify potential malicious use scenarios. Use the results to strengthen security measures and address vulnerabilities. ➜ *Simulating attacks helps uncover weaknesses and improve the model's defenses against real threats.*	🏭 ⛓ 🌐⚡

	Model Deployment	Types of risks mitigated
policy	Strong legal protections and reward programs for whistleblowers Implement strong legal protections and reward programs for whistleblowers who report unethical or malicious use of AI technologies within organizations. ➜ *Protecting whistleblowers encourages the reporting of unethical or unsafe practices, helping to prevent malicious uses of AI.*	☼
technical	Continuously monitor with machine learning techniques Continuously monitor deployed models for signs of intrusion or misuse, using machine learning techniques to detect and respond to threats in real-time. ➜ *Continuous monitoring helps detect and mitigate misuse quickly, preventing significant damage.*	🏭 ⛓ 🌐⚡
technical	Anomaly detection and continuous monitoring in model architecture Incorporate anomaly detection and continuous monitoring mechanisms into model architecture to identify unusual or malicious activities in real-time. Set up alerts for potential misuse. ➜ *Real-time detection and alerts help identify and mitigate malicious activities quickly, preventing significant damage.*	🏭 ⛓ 🌐⚡

Key: Types of Risks Mitigated

⚠	⛓	🏭	🗳	🌐⚡	☼
Human rights abuses	Fraud and crime schemes	Disruption of critical infrastructure	Undermining social cohesion & democratic processes	State conflict	AI risk as a whole

	Model Application	Types of risks mitigated
policy	**Human oversight and control mechanisms** Mandate the inclusion of human oversight and control mechanisms in AI applications, especially for high-risk or sensitive use cases, to prevent fully autonomous malicious actions. ➜ *Human oversight helps contribute to ensuring that AI applications cannot act autonomously in harmful ways.*	⚠ 🌍
technical	**Restrictions on use of foundation models** Foundation model developers may place restrictions on the types of applications that app developers are allowed to utilize foundation models in, and app store providers can do the same for these deployments through app store terms of service or app store regulations. ➜ *Restrictions on application use reduce the risk of AI being used for malicious purposes through widely available commercial services.*	⚠ ⛓
technical	**Red team testing** Conduct red team testing to simulate and identify potential malicious use scenarios. Use the results to strengthen security measures and address vulnerabilities. ➜ *Red teaming helps identify vulnerabilities and improve defenses against real-world attacks.*	🏭 ⛓ 🌍

	User Interaction	Types of risks mitigated
policy	**Legal measures** Utilize existing legal structures to pursue charges against users who utilize AI tools to commit crimes. ➜ *Legal repercussions for misuse act as a deterrent against malicious activities, and relying on existing legislation allows enforcement agencies to act immediately against criminal behavior instead of waiting for specific AI related laws to pass.*	⛓

Key: Types of Risks Mitigated

⚠	⛓	🏭	🗳	🌍	☼
Human rights abuses	Fraud and crime schemes	Disruption of critical infrastructure	Undermining social cohesion & democratic processes	State conflict	AI risk as a whole

	Ongoing Monitoring and Maintenance	Types of risks mitigated
policy	**Reporting mechanisms** Establish clear and widely accessible reporting mechanisms for individuals to report suspected fraud schemes or malicious use of AI technologies. ➔ *Accessible reporting mechanisms help quickly identify and address misuse.*	⛓ 🗳
policy	**Public campaigns** Ensure that reporting mechanisms are well-publicized and easily discoverable through various media outlets, public awareness campaigns, and educational materials. ➔ *Publicizing reporting mechanisms ensures that more people are aware of how to report issues.*	⛓ 🗳
policy	**Confidentiality and protection for reporting** Guarantee confidentiality and protection for individuals who report fraud schemes, ensuring that they are not subject to retaliation or negative consequences. ➔ *Protecting reporters encourages more people to come forward with information about misuse.*	⛓ 🗳
policy	**Regular review** Regularly review and update reporting mechanisms based on user feedback, emerging fraud patterns, and technological advancements to maintain their effectiveness over time. ➔ *Regular updates ensure that the reporting mechanisms remain effective and relevant to changing technological circumstances.*	⛓ 🗳
technical	**Continuously monitor with machine learning techniques** Continuously monitor deployed models for signs of intrusion or misuse, using machine learning techniques to detect and respond to threats in real-time. ➔ *Continuous monitoring helps detect misuse quickly, leading to increased ability to counteract bad actors.*	☼

Key: Types of Risks Mitigated

⚠	⛓	🏭	🗳	🌐	☼
Human rights abuses	Fraud and crime schemes	Disruption of critical infrastructure	Undermining social cohesion & democratic processes	State conflict	AI risk as a whole

Mapping Openness Over the Lifecycle

The Phase I report established a framework for understanding the spectrum of openness in AI model access and the associated risks. The research identified six key risk categories: fueling a race to the bottom, malicious use, capability overhang, compliance failure, taking the human out of the loop, and reinforcing bias. The findings demonstrated that, in general, as access to AI foundation models increases, so does the potential for harm.

Building upon these insights, this section explores how openness affects the AI Lifecycle Framework and risk mitigation strategies associated with it. By mapping the spectrum of openness onto the AI Lifecycle Framework, this report aims to provide a more comprehensive understanding of the unique challenges and opportunities associated with the spectrum of openness at each stage of development and deployment.

This analysis will highlight the critical stages for mitigating model risks throughout the spectrum of openness and discuss how the effectiveness and feasibility of different risk mitigation strategies vary along the spectrum of openness. By synthesizing the findings from both papers, this report seeks to provide actionable insights for policymakers and stakeholders navigating the complex landscape of AI governance in the context of open access models.

Openness and the AI Lifecycle Framework

The degree of openness at each stage of the AI lifecycle can significantly impact the associated risks. For example, during the data collection and preprocessing stage, models with higher levels of openness may benefit from diverse and representative datasets contributed by a wide range of stakeholders, reducing bias risks.

Similarly, the model development and training stage can be greatly influenced by the degree of openness. As access to model components and training methodologies increases along the spectrum of openness, there is greater potential for transparency and collaboration among researchers and developers. This can lead to more robust and reliable models, but it may also make it more challenging to enforce consistent security and ethical standards across all contributors.

The later stages of the AI lifecycle, such as testing and validation, deployment and monitoring, and maintenance and governance, also face unique challenges at different levels of openness. Models with higher levels of openness may require more community-driven approaches, as opposed to the more centralized control possible with models at the closed end of the openness spectrum. For example, ensuring the ongoing security and integrity of a fully open model may rely on a distributed network of contributors and users, rather than a single organization's dedicated maintenance team.

Openness and Risk Mitigations

The effectiveness and feasibility of different risk mitigation strategies identified in the current paper vary significantly along the spectrum of openness defined in the first paper. As the level of openness increases, certain mitigation strategies become more challenging to implement, while others may become more critical in addressing the unique risks associated with models at higher levels of openness.

For example, implementing use case restrictions on fully open models may be significantly more challenging compared to models with lower levels of openness. When a model's components and training methodologies are freely accessible and downloadable, it becomes difficult to enforce limitations on how the model is used or modified by third parties. In contrast, models with lower levels of openness, such as those accessible only through query or modular API access, allow for more centralized control over model usage and modification.

On the other hand, some risk mitigation strategies, such as responsible data collection practices, remain critical across all levels of openness. Regardless of whether a model is fully closed or fully open, ensuring that the data used to train the model is representative of the truth, and free from bias or malicious content is essential for mitigating risks such as reinforcing bias or enabling malicious use.

The risks identified in the first paper, such as malicious use, capability overhang, and compliance failure, intersect with the AI lifecycle stages and risk mitigation strategies in complex ways, and the level of openness plays a significant role in these dynamics. For instance, the risk of malicious use may be higher for models with high levels of openness, as malicious actors have greater access to model components and can more easily modify or fine-tune the model for harmful purposes. In such cases, risk mitigation strategies focused on the model development and training stage, such as secure model architectures and robust testing and validation processes, become increasingly important.

Conclusion

The AI Lifecycle Framework, as detailed in this report, provides a comprehensive and systematic approach to identifying and mitigating risks associated with AI foundation models, with a particular emphasis on the unique challenges and opportunities posed by models further along the openness spectrum. By segmenting the AI development process into manageable stages, this framework enables targeted interventions that are both effective and minimally intrusive, while taking into account the specific considerations surrounding open access development. The alignment with guiding principles ensures that these risk mitigation strategies are feasible, actionable, and balance innovation with risk aversion.

A key focus of the AI Lifecycle Framework is on comprehensive research and deliberate decision-making. By meticulously analyzing every phase of the AI lifecycle, it enables a detailed appreciation of the particular risks and benefits related to AI foundation models, including the distinct implications throughout the openness spectrum. This method, rooted in research, facilitates the pinpointing of specific mitigation strategies based on technical know-how and empirical data, avoiding the pitfalls of vague generalizations or speculative assumptions.

Furthermore, the framework recognizes the necessity of integrating both policy and technical interventions. Policy measures provide overarching guidance and establish regulatory frameworks and standards, while technical mitigations address specific risks at each stage of the AI lifecycle. This dual approach ensures a comprehensive risk mitigation strategy that is adaptable to the evolving landscape of AI technology and the growing prominence of open access models.

The report emphasizes the importance of certain stages in the AI lifecycle for mitigating risks associated with the spectrum of openness, such as data collection and preprocessing, and model development and training. By focusing on these critical stages and implementing targeted risk mitigation strategies, the framework enables policymakers and developers to effectively manage the risks associated with open access AI models while harnessing their potential for innovation and positive impact.

In conclusion, the AI Lifecycle Framework offers a robust and adaptable approach to AI risk mitigation, aligned closely with the guiding principles outlined in this report and tailored to the specific considerations surrounding open access models. By leveraging deep research, targeting both policy and technical interventions, and fostering a culture of transparency, collaboration, and responsibility within the open access AI community, this framework provides policymakers with the tools needed to develop effective, feasible, and actionable strategies for managing the risks and opportunities associated with AI foundation models regardless of their position on the openness spectrum. The collaborative effort of stakeholders across the AI ecosystem, including the open access community, is crucial for ensuring that these strategies are implemented successfully, fostering an environment where open access AI can be developed and utilized safely and ethically.

INSTITUTE FOR SECURITY AND TECHNOLOGY
www.securityandtechnology.org

info@securityandtechnology.org