A Framework for Managing Risks Associated with Artificial Intelligence

Hardly a day goes by without some reference to Artificial Intelligence ("AI") in the news. AI offers great promise across a number of sectors, ranging from healthcare to security, and yet it also carries with it potentially enormous risks, particularly with the advent of Generative AI, which uses existing training data not just to answer questions but to "generate" new content. Leading experts have warned of AI's potentially existential risks, and top executives at Generative AI companies themselves have called for regulation. Companies across industry sectors are confronting questions about how to address these developments. Putting aside the benefits of AI, and focusing first on managing AI-related risks, we need to start by bounding these risks.

AI Risks: What They Are and Why They're Important

AI-related risks[1] tend to fall into three categories: (1) the use of AI technologies as an instrumentality of threat activity, whether from criminals, state actors or disgruntled groups; (2) the malicious targeting of AI systems themselves; and (3) unintended consequences associated with innocent use of AI technologies that can have profound political, economic, ethical and other potentially destabilizing effects.

AI as an Instrumentality of Threat Activity. The first category captures how AI can be weaponized for malicious activity. Some of the risk comes from the ability of AI to increasingly mimic human behavior realistically – a phenomenon using tradecraft known as "**deepfakes**," which are defined as fraudulent "multimedia that have either been created (fully synthetic) or edited (partially synthetic) using some form of machine/deep learning (artificial intelligence)." While synthetic media have been available for decades, Generative AI is enabling the production of highly realistic synthetic content based on much larger datasets. Using this AI-empowered capability, adversaries might do any of the following:

- **Unauthorized access to IT systems.** Generative AI can enable threat actors to develop increasingly realistic social engineering campaigns – not only in text but increasingly voice-related as well. Technology company Retool described in August 2023 how deepfake techniques were used to gain unauthorized access to the company: "The caller claimed to be one of the members of the IT team, and deep faked our employee's actual voice. The voice was familiar with the floor plan of the office, coworkers, and internal processes of the company."
- **Benefits fraud.** Government agencies are starting to see the use of AI technologies to enable fraud in government benefits programs. In July 2023, the Social Security Administration reported that Office of Inspector General agents "discovered that an AI powered "chatbot" was used to impersonate beneficiaries and contact customer service representatives to divert monthly benefit payments to spurious accounts."

[1] In this article, we are principally focused on risks from the use of Generative AI (GAI – e.g., ChatGPT, MidJourney, Stable Diffusion, Dall-E, etc.) and other narrow AI systems such as expert learning models. We do not address risks from Artificial General Intelligence, whose development is rather further in the future.

- **Influence and Disinformation.** It will also be important to garner a better understanding of how Generative AI could potentially be used as a tool to spread false or misleading information that could, for example, affect the company's reputation by eroding consumer trust. As described in an April 2023 Foreign Affairs article, language models, dubbed "personalized propaganda," have already been trained to persuade game players to partner with them in a game, and they could be readily trained to persuade people to take actions with real world commercial effect, to include changing product preferences due to a negative image. Likewise, in March 2022, a deepfake video was released of Ukrainian President Volodymyr Zelensky appearing to announce a surrender to Russia.

Not all AI threats rely on human mimicry. Indeed, some of the more significant threats of AI weaponization arise directly from its enhanced capabilities and include, for example:

- **Advanced malware.** Researchers have demonstrated how platforms like ChatGPT could be used to create polymorphic malware (which mutates to change its code while retaining its core function).
- **Surveillance.** A team at Carnegie Mellon used AI systems to pinpoint locations of humans using only the interference caused by WiFi signals in a room, thereby turning every WiFi router into a potential surveillance device.
- **Chemical Biological Radiological and Nuclear (CBRN) threats.** Security experts have warned how artificial intelligence could be potentially misused for the creation of new catastrophic biological and chemical weapons, as in this article by former DHS official Paul Rosenzweig. An October 2023 RAND Corporation study on potential AI misuse for weapons development found that, while large language models had not generated explicit instructions for creating biological weapons, they did offer guidance that could assist in the planning and execution of a biological attack.

Generative AI platform providers like OpenAI and Google have released prohibited-use policies, which are enforced through business logic on the platforms themselves, but threat actors are already using prompt engineering techniques to bypass such policies. This will inevitably lead to a cycle of litigation, regulation and more intensive content moderation programs, which will engender many of the same challenges (and more) faced by social media platforms in enforcing content moderation policies on their platforms. As open-source AI evolves, Generative AI capabilities will likely increasingly be in the hands of organizations with limited to no content moderation policies or those with mal-intent who are difficult to deter.

Malicious Targeting of AI Systems. Regarding the second category (malicious targeting of AI systems themselves), as a company's dependencies on AI systems increase, these systems will be increasingly targeted by threat actors – whether to illicitly obtain sensitive information, disrupt core business functions for extortion purposes, or make a political statement. Inasmuch as AI is based in software code, the entire panoply of cybersecurity risk applies with equal force (if not greater) to AI systems on which an enterprise might depend. These risks also highlight the importance of **supply chain risk management**, particularly where organizations are sourcing AI requirements to third party AI technology providers.

Organizations need to be on the lookout for malicious prompt injections (basically hijacking). A prompt injection essentially tells the model to ignore previous instructions and do something else – for example injecting malicious code writing instructions, impacting AI bot assistants, and the like. On a related note, researchers have highlighted how prompt engineering can be misused to obtain information about previous prompts, API connections and even data supporting third party AI applications.

Similarly, where data is crowdsourced, companies can also be subject to data mischief-making, sometimes called data poisoning. One of the best known, if slightly dated, examples, was the one-day existence of a Microsoft Twitter bot named "Tay," which relied on its scanning of a large volume of Twitter posts to generate its own responses to Tweets. Perhaps unsurprisingly, within a matter of hours the bot had turned into a racist jerk and had to be shutdown. Likewise, in 2017 researchers used stickers to trick self-driving cars into ignoring stop signs. Indeed one of the voluntary commitments made by leading AI companies in July 2023, and described in a White House announcement ("the White House 2023 Voluntary AI Commitments") is a promise to "invest in cybersecurity and insider threat safeguards to protect proprietary and unreleased model weights," which are essential elements of any AI model.

Then too, malicious actors may simply disrupt an AI system and make it unavailable, as reflected in Fall 2023 distributed denial of service (DDoS) attacks against leading AI platforms. If an enterprise's business model depends, for example, on AI generated customer service response, the potential for adverse malicious action is palpable.

Unintended Consequences. As for the third category, AI systems could increase the potential for unintended consequences, both for inputs and outputs related to this technology. Input-related risks include both leaks of sensitive data being loaded into prompts and underlying models, as well as the quality of prompts themselves. There have been a number of instances where engineers, software developers and other company employees have unintentionally leaked sensitive corporate information such as software source code to third party Generative AI platforms. More generally, prompts are what guide a Generative AI model's output. Poor quality prompts can confuse the model and yield poor quality results.

Outputs represent the scarier unintended risk. Generative AI systems rely on large volumes of data to build their models, and as with previous generations of AI-based systems, it is "garbage in, garbage out." Generative AI systems face these same problems on a much larger scale. They are likely to reflect the explicit (clear and obvious) and implicit (unconscious or unrecognized) biases contained in the source data on which they were trained, particularly when the data is not carefully curated. Perhaps more troubling, flawed data sampling can reinforce existing biases and outcome gaps by under-sampling minority and marginalized groups, creating higher error rates and worse outcomes for those groups, as has been seen in many facial recognition technologies. These are difficult issues without easy answers.

As another example of output risk, hallucinations are absurd conclusions generated by AI applications based on Generative AI algorithmic flaws, poorly constructed prompts or bad data. They could drive error-prone automated decision-making within any functions using Generative AI tools. Also, when an expert system AI model is given a task that it executes autonomously, it can be tricked into making unintended decisions. Consider the example cited above where researchers tricked self-driving cars into ignoring stop signs. This level of autonomy could create challenges when it comes to assigning accountability for any negative outcomes resulting from the actions of AI systems. When an AI system makes a decision or takes an autonomous action that results in harm, who is accountable? The user of the technology, or its developer? AI systems are astonishingly complex, with the largest systems reportedly exceeding one trillion parameters (that is, inputs or variables that help determine the output of a model).

AI Risks: What to Do About Them

Defending Against the Weaponization of AI. Having framed these risks, how should leaders manage them? The first category (use of AI technologies as an instrumentality of threat activity) tends to fall more squarely in the domain of security, fraud and crisis management teams. Managing these types of AI risks benefits from a threat-informed defense operating model where threat intelligence and security engineering are fused to understand how AI may tend to weaken the effectiveness of certain security technologies – for example email filtering or endpoint protection systems – and what countermeasures need to be implemented to address them (e.g., reputational analysis, more user-centric behavioral analytics). Communication teams run exercises on developing situational awareness for, and responding to, disinformation campaigns.

A broader concern exists regarding advanced AI technology falling into the wrong hands or being otherwise misused. In October 2023, President Biden signed an Executive Order (EO) on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence . The EO directs the Department of Commerce to require significant additional reporting from the developers, operators and providers of significant foundational AI systems (referred to as “dual use foundation models” and “large scale computing clusters”). It also calls for steps to develop best practices and standards, and it puts a special focus on steps that might be needed to mitigate CBRN and cybersecurity-related risk.

Defending Against Malicious Targeting of AI. The second category (malicious targeting of AI systems) also relies on security teams, but there is a significant dependency on business and CTO functions to gain visibility on AI platforms used within the organization, and related use cases, so they can be appropriately protected. With this visibility in hand, organizations can start to think about whether AI models are protected behind secure control gates to prevent bad actors from stealing or otherwise misusing these models. MITRE’s ATLAS platform and NIST’s white paper on adversarial machine learning can help to identify risks and plan defenses, and a set of ATLAS mitigations are now available in draft.

That, for example, is why the EO requires certain developers to report on the physical and cybersecurity protections taken to assure the integrity of the AI product.

Managing Risks Regarding Unintended Consequences of AI Usage. For the third category (unintended AI consequences) to be effectively managed, business and technology teams need to lead. Not only do organizations need to address operational and reputational risks associated with unintended consequences, but the regulatory and litigation environment is rapidly shifting and requires constant monitoring. In June, the European Parliament (EU) approved draft AI legislation that would restrict the use of facial recognition software, prohibit the use of AI for "social scoring," scrutinize the use of AI in support of critical infrastructure operations, and require disclosure around underlying data sources and the logic used by algorithms (otherwise known as the "right to explain").

California has already enacted legislation that requires AI bots used in online interactions to identify themselves as such. And U.S. Federal Trade Commission (FTC) Chair Lina Khan penned a May 2023 New York Times Op Ed intimating that the Commission will use existing authorities such as the Equal Credit Opportunity Act to pursue actions against AI products that result in what it perceives as discriminatory outcomes. On the litigation front, a class action lawsuit was recently filed against OpenAI (the creator of ChatGPT) and Microsoft alleging a series of privacy infringements, deceptive trade practices and unfair competitive practices. As noted above, in July 2023 the White House announced a set of voluntary AI commitments by leading AI companies around safety, security and trust, followed by the October 2023 EO.

In the short-term, across the companies we work with who are simply consumers of AI technology, we are observing trends around a three-pronged approach: (1) gain visibility on actual AI use; (2) apply risk-based policies; and (3) educate the workforce on AI-related risks. To support these activities, companies are also using existing Information Sharing & Analysis Organizations (ISAOs) to share information on AI risks and best practices.

For organizations actively using AI technologies, consider how to implement guardrails around data, disclosure and decision-making – the "Three Ds."

- On data, the focus should be around both accuracy and whether they have the legal right to use the data in question – i.e., was it aggregated without the consent of the data owner, or does data fall under a "fair use" exception. Fair use is a legal doctrine that permits unlicensed use of copyright-protected works in certain circumstances, such as criticism, comment, news reporting, teaching, scholarship, and research.
- On disclosure, where organizations are displaying text or video content developed by AI technologies, they should disclose that fact. One of the White House 2023 Voluntary AI Commitments is for leading AI companies to develop provenance and/or watermarking systems to determine if a particular piece of content was created with their system, and the EO also includes synthetic content disclosure requirements.
- On decision-making, organizations should have in place human review for any important AI-informed decision that affects a human being, particularly for outcomes that could potentially have a disproportionate impact on protected classes of individuals. The EO requires, for example, that specified agencies administering Federal benefits to ensure

appropriate human review of AI decisions or that denial-of-benefits appeals go to human reviewers.

It is equally important to have human review when AI systems take critical decisions that might substantially impact an enterprise's vital business functions. Though routinized AI-aided decision making is certainly the wave of the future, responsible governance will require human input on essential decisions for some time to come.

In the longer term, as AI advances, it will be important for organizations to implement a formalized governance and accountability framework around the use of increasingly powerful AI systems. Someone needs to be accountable for making sure that the models being used are appropriate for the task, have "boundaries" in place to prevent their misuse, and are stress-tested to discover and address unintended consequences in a timely manner.

In January 2023, the U.S. National Institute of Standards and Technology (NIST) released an AI Risk Management Framework, which classifies key risks associated with the use of AI and defines structures for framing, governing, mapping, measuring and managing the use of AI inside organizations, including through test, evaluation, verification, and validation (TEVV) processes.

This NIST Framework establishes the concept of “trustworthy AI” and defines characteristics of trustworthy AI systems, such as being valid and reliable, safe, secure and resilient, accountable and transparent, explainable and interpretable, privacy-enhanced, and “fair,” meaning that the potential for harmful bias is managed. The EO directs NIST to develop a companion framework addressing Generative AI-specific risks.

Industry organizations are stepping forward with industry-specific principles that support AI-related governance and strategic planning. In November 2023, the National Retail Federation released its Principles for the Use of Artificial Intelligence in the Retail Sector, which provide guidance on governance and risk management, customer engagement and trust, workforce applications, and business partner accountability.

Of course, there won’t always be clear answers when applying these factors. Some of the key choices in AI design are akin to earlier technology tradeoffs. For example, the tradeoffs of using open-source AI models such as Meta’s LlaMa versus closed-source models such as OpenAI’s ChatGPT have parallels to using open-source versus proprietary software libraries, with the former offering the benefits of crowdsourced-improvements and transparency but suffering from uneven levels of maintenance and contributor guardrails. Likewise, the choice of using large parameter general purpose models versus special purpose models with fewer parameters are akin to using general purpose computers (like PCs) versus special purpose machines (think of an IOT devices or embedded operational technology systems).

These differences will have an impact on the trustworthiness of the AI under consideration. Closed systems, for example, may be more secure while open systems will be more transparent and explainable. Finding the sweet spot for AI risk/reward tradeoffs will be highly context and enterprise specific.

As with software development, many companies will also face a tension between time-to-market pressures for quickly deploying AI technologies to gain competitive advantage versus slower and more deliberate approaches to ensure greater AI trustworthiness (plus compliance with applicable EO provisions). Ideally, companies would start by limiting the use of AI to internal functions, so they can become familiar with AI technologies, before offering AI-driven services to customers and business partners. More narrowly tailored use cases – for example use of facial recognition-related AI for one-to-one matching versus one-to-many surveillance use cases – generally entail significantly lower risks. Strong governance can also help ensure that thoughtful planning, resourcing, and validation are applied to addressing these questions. Borrowing from secure software lifecycle best practices, the White House 2023 Voluntary AI Commitments also include promises to implement (a) internal and external red-teaming of AI models as well as (b) bounty systems to incent the responsible disclosure of AI weaknesses and unsafe behaviors (or to include AI systems in their existing bug bounty programs). The EO also applies red team testing more broadly, and also directs NIST to develop a companion resource to the Secure Software Development Framework to incorporate secure development practices for Generative AI and for significant foundational models.

To conclude, AI represents a rapidly advancing technology, and organizations are going to need to think differently about how to defend against AI-related threats and risks. Organizations should consider whether defensive architectures need to be modified and tested for AI-enabled threats such as advanced social engineering or new forms of malware. They should also reflect how to prepare for disinformation campaigns directed at customers, employees and other stakeholders. For organizations actively using AI technologies, they should prepare for scenarios where the data or algorithms underlying these technologies is subverted. Finally, they should prepare for unintended consequences related to the innocent use of such technologies, specifically by putting guardrails around the "Three Ds" – what data is used, how any use of AI technology is appropriately disclosed, and how any AI-enabled decision-making includes a human in the loop.

About the Chertoff Group.
The Chertoff Group is an advisory firm of highly qualified experts that uses proven frameworks to help organizations achieve their business and security objectives in a complex risk environment. Our team helps organizations manage cyber, physical and geopolitical risks; navigate evolving regulatory and compliance requirements; and discover opportunities to win business and create value. Through our investment banking subsidiary Chertoff Capital, the firm provides M&A advisory services to companies in the defense technology, national security and cybersecurity markets. Together, we enable a more secure world. For more information visit www.chertoffgroup.com or contact info@chertoffgroup.com