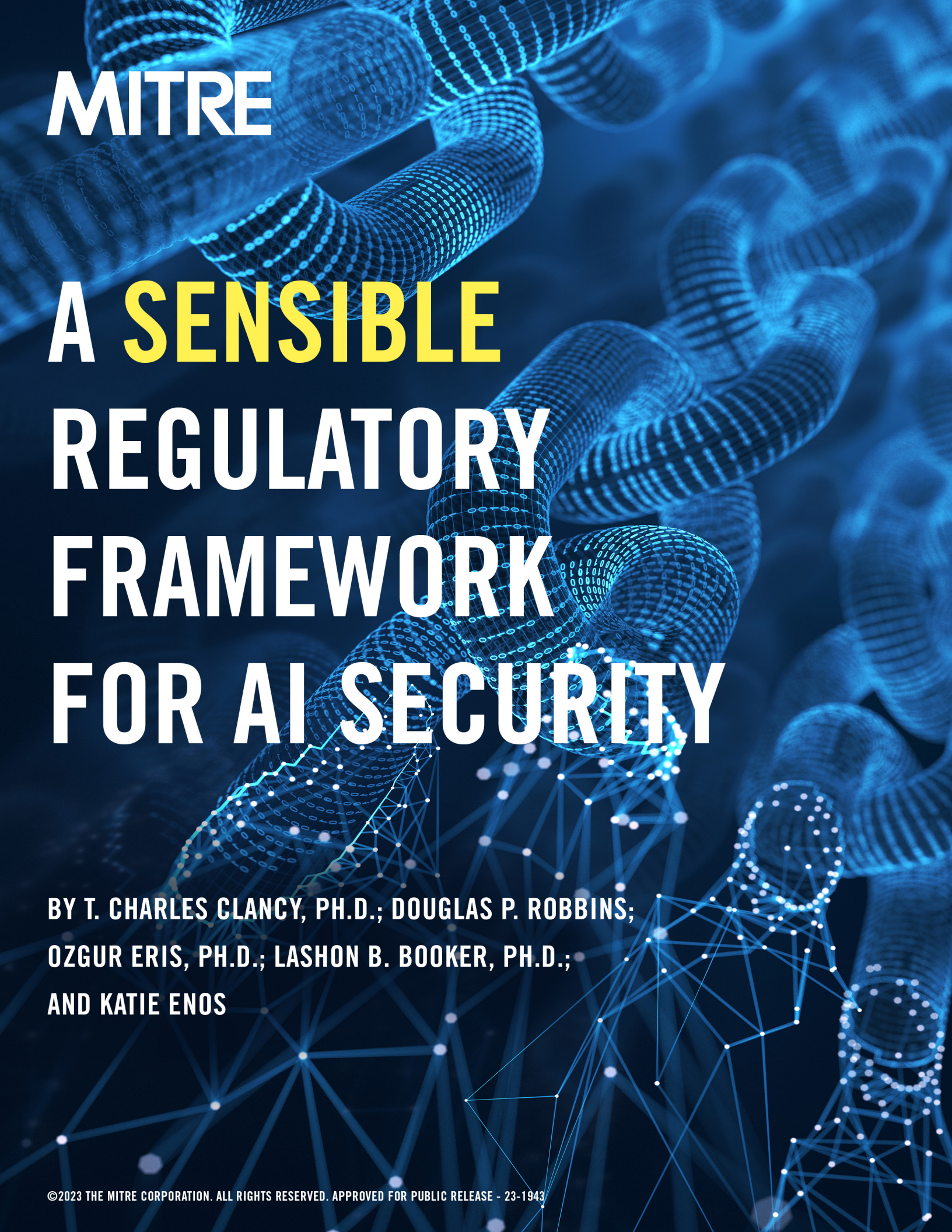

MITRE

A SENSIBLE REGULATORY FRAMEWORK FOR AI SECURITY

BY T. CHARLES CLANCY, PH.D.; DOUGLAS P. ROBBINS; OZGUR ERIS, PH.D.; LASHON B. BOOKER, PH.D.; AND KATIE ENOS

Contents

Background

Over the past ten years, and more visibly over the past six months, artificial intelligence (AI) has gone through tremendous technological advancement. In 2012, advances in graphics processing units (GPUs) enabled the first example of tractably training a deep neural network to outperform more traditional approaches to machine learning. In 2017, Google researchers published their paper on transformer networks that provided the building blocks for large language models (LLMs). These breakthroughs, along with many other innovations, resulted in machine capabilities such as:

- Improved machine perception that for certain tasks can exceed human cognitive visual performance.
- Optimization and planning engines that leverage reinforcement learning to exceed human performance in complex games.
- Generative algorithms that can create text, audio, and images that in many cases are indistinguishable from those created by humans without targeted analysis.

Collectively, these advancements lead many to believe that artificial general intelligence (AGI) is just around the corner. At the same time, AI is increasingly being recognized as a revolutionary technology that can aid the government in addressing critical, mission-oriented challenges, ranging from healthcare to national security.

With any new, disruptive technology, humanity looks at how to shape its development and application. With AI, these discussions range from geopolitical—how AI affects the balance of power between the United States and China, to existential—that a super-intelligent AI may be a threat to humanity itself.

Given the exponential pace of technological advancement, both excitement and anxiety about the formation of superintelligence are endemic across the tech and policy ecosystems.

Within this paper, we explore potential options for AI regulation and make recommendations on how to establish guardrails to shape the development and use of AI.

AI Threats and Risks

Any attempt to secure or regulate a new technology should be informed by its **vulnerabilities, threats** that exploit those vulnerabilities either intentionally or unintentionally, and the ultimate **risk** of damage, harm, or loss to human life, health, property, or the environment. This conceptualization facilitates effective threat modeling and risk management.

The national dialogue often lumps together a wide range of algorithmic technologies under AI, which can confuse this analysis. Here, we divide the AI ecosystem into three broad categories based on how AI can be manifest in application:

1. **Engineered systems that use AI as a component or subsystem**
2. **AI as an augmentation of human capabilities**
3. **AI operating autonomously under its own agency**

Differentiating these categories is important because the threats and risks differ based on how AI manifests in applications, as do the approaches to mitigating threats and risks.

AI as a Subsystem

AI is embedded in many software systems. Discrete AI models routinely perform machine perception and optimization functions, from face recognition in photos uploaded to the cloud, to dynamically allocating and optimizing network resources in 5G wireless networks.

There are a wide range of vulnerabilities and threats against these types of AI subsystems—from data poisoning attacks to adversarial input attacks—that can be used to manipulate subsystems, with the goal of having a deterministic and malicious impact on the

target system. While the overall software system has its own vulnerabilities, those vulnerabilities can now be mostly understood through traditional test and evaluation, validation and verification, and cybersecurity lenses. The introduction of an AI subsystem may introduce new, unknown, and unique vulnerabilities that are largely unexplored at this point. Tools are emerging to identify and protect against new and existing threats. For example, MITRE works closely with industry and government to capture such threats and document associated adversary tactics, techniques, and procedures in the MITRE Adversarial Threat Landscape for Artificial-Intelligence Systems (ATLAS)™ framework.[1] Building on ATLAS and in partnership with Microsoft, MITRE released tools to perform red team testing of converged AI-cyber systems as Arsenal.[2] Meanwhile, a decade of research into AI assurance is now turning into robust industry best practices such as **model cards** that establish the boundaries of an AI model's use and can help inform developers, policymakers, ethicists, and users.[3]

MITRE defines AI assurance as a lifecycle process that provides justified confidence in an AI system's ability to operate effectively with acceptable levels of risk to its stakeholders. An AI system operating effectively means that it meets its functional requirements with valid outputs. The risks that need to be managed within acceptable levels may be associated with or stem from a variety of factors depending on the use context, including but not limited to AI system safety, security, equity, reliability, interpretability, robustness, privacy, and governability.

As this definition suggests, assuring AI necessitates that digital systems with AI-based components and subsystems have a full lifecycle of scrutiny. The National Institute of Standards and Technology's (NIST) AI Risk Management Framework (RMF)[4] is a good example of an approach that incorporates trustworthiness considerations into AI design, development, use, and evaluation of AI system components.

While many open questions remain in this domain, such as user data privacy rights, AI assurance researchers and practitioners have useful framings of the problem to mitigate threats and manage risks.

AI as Human Augmentation

Another application of AI is in augmenting human performance, allowing a person to operate with much larger scope and scale. This has wide-ranging implications for workforce planning as AI has the potential to increase productivity and shift the composition of labor markets, similar to the role of automation in the manufacturing industry.

LLMs are front and center in this debate, as they demonstrate human-level performance in myriad white collar job tasks. However, a growing concern is that bad actors are augmenting their adversarial capabilities by leveraging AI, especially in cyber operations and mis/disinformation.

For cyber operations, LLMs could be immediately useful in identifying vulnerabilities in computer systems, from software bugs to configuration errors. Additionally, they could make an efficient "co-pilot" in planning and executing cyber operations by synthesizing large amounts of network metadata and using that context to efficiently propose courses of action weighed by the likelihood of success in exploiting a computer network. These types of systems could emerge and be used in active cyber conflict before the end of 2023.

For mis/disinformation, the generative nature of LLMs also makes them well-suited for generating high-quality content quickly, from phishing email campaigns to nation-state propaganda. While sophisticated hackers and military information operations can already generate believable content today using techniques such as computer-generated imagery, LLMs will make that capability available to anyone, while increasing the scope and scale at which the professionals can operate.

While neither cyber attacks nor the spread of mis/disinformation represent new classes of activity, disruptive AI technologies, such as LLMs, will likely cause their further proliferation.

AI with Agency

A segment of the tech community is increasingly concerned about scenarios where sophisticated AI could operate as an independent, goal-seeking agent. While science fiction historically embodied this AI in anthropomorphic robots, the AI we have today is principally confined to digital and virtual domains.

One scenario is an AI model given a specific adversarial agenda. Stuxnet is perhaps an early example of sophisticated, AI-fueled, goal-seeking malware with an arsenal of zero-day attacks that ended up escaping onto the internet.[5] While no LLMs were involved, and its autonomy logic was modest by modern standards, it is nonetheless an example of autonomous malware.

Perhaps the recent equivalent of this is Auto-GPT, where a user can express a goal in natural language, and it uses GPT-4 to divide that task into sub-tasks and uses the internet to accomplish them.[6] While some are using Auto-GPT for tasks like personal finance optimization, Chaos-GPT is an example of an instance with the potential for nefarious intent and the goal of exterminating humanity.[7]

Another theorized scenario is that an Auto-GPT–like system might independently develop a sub-task that prioritizes harming humanity.[8] Ironically, AI developing such a sub-task would most likely be based on or inspired by human-generated content that serves as input to the AI, since AI does not have any inherent goal or purpose. While many dismiss such a concern as hypothetical science fiction, it may not matter if people develop things like Chaos-GPT intentionally.

However, an AI application like Chaos-GPT still interacts with the world through the same interfaces as humans. Anything it can do over the internet, a rogue nation-state could theoretically also do with human fingers on a keyboard, though the AI and associated automation accelerates timelines and increases the scale of the threat. Therefore, it is even more essential to make sure the safety-critical cyber-physical systems that control our environment are secure against exploitation. Because whether it's a human, an AI-augmented human, or a malicious AI agent, they all rely on the same tactics, techniques, and procedures to exploit our infrastructure.

Regulatory Approaches

A debate about how to regulate AI has captured the nation's attention, and there is no shortage of approaches attempting to address concerns raised in each of the three categories outlined above: AI as a subsystem, AI as human augmentation, and AI with agency.

In April, the Future of Life Institute published a set of recommendations for policymaking that are representative of many of the concepts being considered:[9]

1. **"Mandate robust third-party auditing and certification.**
2. **Regulate access to computational power.**
3. **Establish capable AI agencies at the national level.**
4. **Establish liability for AI-caused harms.**
5. **Introduce measures to prevent and track AI model leaks.**
6. **Expand technical AI safety research funding.**
7. **Develop standards for identifying and managing AI-generated content and recommendations."**

Lawmakers are also debating concepts such as including protections for personal data and building transparency guidelines so that there is clear

accounting for what data is being used to train AI models; establishing common definitions for AI models, applications, and capabilities; strengthening the AI workforce; and licensing AI models.

Some of these regulatory concepts have merit but others may limit competition and advancement, weaken international leadership, or establish regulatory controls without necessary sector context. The multi-million-dollar cost to train an LLM from scratch, restricting research and development[10] primarily to the major tech companies and well-funded startups, like OpenAI, surfaces additional dilemmas for policymakers:

- Big tech companies have a monopoly on the future of AI, with their large budgets preventing startups and competitors from advancing as quickly, and potentially giving a small number of companies a disproportionate influence on AI-generated content.
- Efforts to regulate AI could be meaningfully advanced by engaging with only a handful of key tech companies, such as attendees of the White House meeting in early May.[11]
- Competition between China and the United States on AI can be meaningfully measured by comparing the progress of U.S. hyperscalers, service providers that offer massive, elastic computing resources, to that of Chinese hyperscalers.

MITRE would discourage focusing regulations only on big tech, as open-source innovation in AI has enabled smaller companies to have greater access to AI tools and technologies.

Furthermore, AI is on a short list of technologies being competed internationally, with outcomes over the next several years defining how well individual countries will be able to protect their national and economic securities. The United States cannot cede international leadership on this important technology. Instead, we must find a way to navigate our interest in advancing a new generation of technologies within an appropriate regulatory framework.

Illustrating that point, many have latched onto the second recommendation above, which seeks to regulate access to computational power. Producing and using GPUs is more visible and feasible by fewer state and non-state actors than enriching uranium,[12] inciting an international arms control regime around GPUs.[13] Presuming international treaties around such an arms control regime are agreed upon, then members of the international community should be prepared to use all levers at their disposal, including kinetic force, to prevent nations from training too sophisticated of an AI model.[14]

The major challenge with compute as the regulatory throttle is that we have already seen innovation that reduces reliance on compute, so this also appears to be a misguided approach to regulation.[15] Moreover, such a regulatory mechanism would be exceedingly challenging to implement on a global scale, and if only partially implemented, could give our adversaries a competitive advantage and adversely affect our security.

Congressional hearings on this topic in May surfaced the challenges of developing actionable policy to address these concepts. A variety of factors concerning regulation have been considered in previous studies,[16] and these factors suggest the key questions to be addressed in an analysis:

- **What are the objects to be regulated?** There is no single, widely accepted definition of AI, and there are many choices for what to regulate (e.g., data, algorithms, testing, market access).
- **What are the problems regulation is intended to address?** Concerns can range from considerations like accountability, discrimination, and privacy to technical issues like dependability, transparency, and accuracy.

- **What is the most appropriate governance mechanism?** Proposals have been offered, ranging from voluntary self-regulation to government-mandated policies and procedures.
- **What concrete regulatory instruments are appropriate for achieving which purposes?** Regulatory actions have been discussed for safety and accountability (bans, approvals, standards for design, liability standards, third-party auditing and certification, etc.), for transparency (disclosure of technical details, explanations for decisions, disclosure of where AI is used), and for the involvement of stakeholders (researchers, developers, users).

Proposals to establish a single federal agency focused on regulating and licensing AI miss the fact that AI is, or soon will be, a fundamental element to most aspects of technology. It would be more scalable and less duplicative to focus AI regulation on the point at which AI intersects a regulated industry. For example, rather than a new agency regulating AI in general, the Food and Drug Administration should extend its current "Software as a Medical Device" regime to cover AI-enabled software in a healthcare setting. While these existing regulators will all need to beef up their AI expertise, they can leverage their existing deep domain expertise to better contextualize regulation and licensing.

The three operational categories discussed on page 3 will have unique regulatory considerations. This is further discussed in the next section.

Proposed Elements of a Regulatory Framework for AI Security

This paper's focus is primarily on AI security, and hence we propose elements of a regulatory framework based on a vulnerability, threat, and risk calculus. Risks are realized when threat actors (intentional or not) exploit vulnerabilities. We intersect this calculus with our AI technology space decomposition of AI as a subsystem, AI augmenting humans, and AI with agency. While there are several actions that need to be taken across this matrix, the table below highlights the three that are most critical for immediate action.

	AI Vulnerabilities	AI Threats	AI Risks
AI as a Subsystem	Reduce vulnerabilities by enhancing industry-specific approaches		
AI Augmenting Humans		Limit threat and hence risk scaling through human penalty, supported by increased auditability	
AI with Agency			Reduce risks via critical infrastructure hardening and enable safe research

Regulatory considerations for AI as a component or subsystem

1 **Any AI regulation should require AI components to satisfy software assurance requirements as well as AI-specific assurance requirements that can be developed based on validated AI assurance frameworks.**

When considering AI capabilities embedded as a component in some larger engineered system, it is important to first recognize that the AI component is a piece of software. Whatever software engineering best practices and regulatory requirements are enforced for the larger system should also be enforced for the AI component. This can sometimes be a challenge, though, since AI software exposes new vulnerabilities that differ from traditional software risks (e.g., lack of testing standards, lack of widely accepted code development practices, reliance on training data and other external information influencing program behavior).

2 Regulated industries should develop a NIST AI RMF response plan; if they deem the NIST AI RMF insufficient, they should identify alternative AI assurance approaches.

The NIST AI RMF shows how to systematically think about risk during the design, development, and deployment of an AI system. Though compliance with the NIST framework is voluntary, it is an important first step toward identifying where regulatory mandates might eventually be required. Regulated industries, particularly those involving higher risks, should develop a response that outlines how they intend to apply the NIST framework to their sector.

3 Any AI regulation should account for and mitigate risks stemming from component interactions.

Many AI components are designed and implemented as integrated systems that leverage several AI technologies, along with more conventional methods, to perform a computation. One of the lessons learned from experiences with safety-critical software applications[17] is that accidents and failures involving complex software systems are much more likely to result from dysfunctional interactions among components than from individual component failure. The interactions and tight coupling among subcomponents in an AI system can be a source of vulnerability to threats and risks for use/misuse that may require management and regulation. Consequently, it may be prudent to view component interactions as regulatory objects subject to testing standards and assurance management requirements.

4 Any AI regulation should require "assurance cases" to be developed before deployment.

Another insight from the safety-critical software community is that best practices, technology improvements, standards, and testing are all necessary but not sufficient to provide adequate assurance guarantees. It is important to rigorously tie all these factors together into "assurance cases"[18,19] that assess the risks of failures and harmful outcomes before a system is deployed. An assurance case is a documented body of evidence that provides a compelling argument that the system satisfies certain critical assurance properties in specific contexts. It includes explicit claims about critical system assurance properties and behavior boundaries, evidence for these claims, and a rigorous argument that demonstrates that the evidence is sufficient to establish the validity of the claims.

AI systems have specialized concerns from an assurance perspective. For example, an assurance case for a machine learning system might describe the intended performance envelope of the system for a given input distribution and identify mechanisms to detect violations of the data assumptions and issue alerts. Note that an assurance case for an AI system would go beyond the information available in something like a model card, since the assurance case would use the evidence to build an argument focused on convincing stakeholders that an assurance claim is true.

5 Any AI regulation should account for use context and favor existing domain-specific regulations.

For regulation, the context in which an AI component is used is as important as the aforementioned software perspectives. Existing regulated sectors or industries consider the deployment context for an application. Given this, when an AI-enabled application is deployed in a regulated sector or industry, AI regulation should first be addressed through existing regulatory frameworks in that domain. Many issues regarding performance standards, oversight, and legal responsibilities can be handled with existing regulations and laws, though some maturation of requirements may be needed to deal with particular concerns of AI-enabled components (e.g., transparency, post-deployment monitoring) There is a need for regulatory flexibility to accommodate a rapidly changing technology where innovations continually challenge expectations about what is possible.

6 Industry regulators should conduct continuous regulatory analysis of individual use cases.

Recommendations for regulatory action will change as aspects of the technology and use case change over time. One way to address this challenge is to conduct a regulatory analysis of the factors influencing regulatory decisions and their interdependencies, then use that analysis to organize policy questions and craft appropriate regulations to match the properties of each use case. This will help avoid regulatory actions that take an inflexible one-size-fits-all approach. Those processes should also explore how emerging AI solutions impact citizens, commerce, and government and identify— in parallel with AI solution development— associated governance needs such as defining application-specific assurance requirements and overseeing compliance. As the AI assurance regulatory space is populated, agencies can continuously monitor and bridge emerging disconnects between existing AI governance and the risks introduced by new AI solutions through rapid adaptations. This is particularly relevant for a rapidly developing technology such as AI and may require AI assistance to accomplish.

7 Industry regulators should promote trusted information sharing mechanisms to support regulatory analysis.

Even a simple analytic step like establishing a common vocabulary that links use case attributes to harmful outcomes and effective mitigations can provide important leverage for assurance. For example, MITRE's ATT&CK® framework, a globally-accessible knowledge base, describes cyber threat use cases using carefully selected attributes that make it possible to link use cases with appropriate response actions in a systematic manner. MITRE's ATLAS framework extends that type of aggregation to AI systems. Such information sharing facilitates pattern detection across use case types and identification of plausible candidate actions for new use cases involving technology innovations. That body of knowledge can potentially form the basis for crafting regulatory actions like certifications for systems claiming to provide certain kinds of cyber defense capabilities. The Federal Aviation Administration's Aviation Safety Information Analysis and Sharing, or ASIAS, data sharing initiative in the aviation industry may also provide guidelines for how to proceed.

Regulatory considerations for AI implementations that aim to augment human capabilities

8 AI regulation should require system auditability in order to hold individuals who misuse AI to cause harm accountable.

When the augmented human is a bad actor with ill intent, the potential harms that bad actor can realize are amplified. This scaling of the threat, and hence risk, is much more tangible and plausible in the near term than the risk of AI rapidly evolving into AGI taking on systematic ill intent toward its surroundings and the human race. This risk also essentially originates from humans and the intent to do harm—defrauding, cyber hacking for financial or other gain, or causing societal uncertainty through social media influence. It should also be noted that augmenting humans with AI can have a positive effect. For example, increasing a pathologist's ability to find cancer cells through AI augmentation has life-saving promise. Given these factors, it is not particularly feasible or desirable to either limit use of AI in augmenting humans or to attempt to hold AI accountable for augmenting a human; it would be more effective to penalize and deter bad actors with ill intent. Our legal code primarily relies on holding humans causing harm with ill intent accountable. A key aspect of related regulation should focus on enabling system auditability to document the intent of humans using AI to cause harm and the execution of that ill intent with AI. Without such auditability, it would be challenging to argue for accountability, and, thus, to establish legal frameworks for deterrence.

9 **Legal frameworks to deter intentional and harmful AI misuse should scale accountability with risk.**

Legal frameworks that may be developed as part of AI regulation to hold individuals accountable for causing harm with AI should focus on regulatory approaches commensurate with the scale of the risk. For example, increased legal penalties should be tied to higher degrees of harm caused by intentional AI misuse. Moreover, regulation should ensure that our legal code sufficiently articulates the harms and differentiates between intentional acts and "AI accidents," a concept that will evolve as AI technologies advance.

10 **AI regulation should provide appropriate levels of transparency into AI applications to an objective third party and/or the public for detection and mitigation of intentional AI misuse.**

As discussed on page 4, we view the emerging LLM ability to generate realistic, high-quality content as having strong potential to sow mis/disinformation and adversely influence highly consequential activities like a national election. While increasing auditability provides a legal means for accountability and hence penalty and deterrence, those actions lag harms, and today's technologies may enable various degrees of anonymity despite auditability. This calls for additional regulatory approaches that enable near-real-time interventions to mitigate risk. Regulation could focus on enabling third-party watchdog functions by requiring solution providers to provide real-time means to transparently monitor trends and, potentially, content within privacy bounds. This could include the public through tools similar to MITRE's SQUINT™, a browser plugin and mobile app focused on allowing the public to spot and mitigate COVID-19 mis/disinformation. Additionally, the ability to automatically detect fake content (prose, images, and videos) is nascent and requires national research and development investment.

Regulatory considerations for AI implementations that have agency

We expect that risks from a sophisticated AI operating with agency and malintent are less likely in the near term than those introduced by AI augmenting human actors. Regardless, in both cases, the human or AI actor must use the same "hands on keyboard" mechanisms a rogue nation-state would employ. These mechanisms are either intentionally open interfaces or more likely vulnerabilities caused by poorly designed or implemented systems that expose interfaces that can be actuated to cause harm. Of particular concern are safety-critical cyber-physical systems—those that through a cyber-physical interface create the opportunity for injury or death to people, the loss or damage of equipment or property, or environmental harm. Given the speed at which an AI might operate, the threat and risks are amplified; hence, additional attention is required to accelerate mitigations to counter.

11 **Federal government critical infrastructure plans should address increased risk due to AI-enabled scale and speed and consider countering risk with automated red teaming.**

We recommend an assessment of federal government critical infrastructure plans focused on identifying and strengthening recommendations for safety-critical cyber-physical systems particularly vulnerable to increased threats due to the scale and speed AI enables. Additionally, we recommend sector-specific use of advancements in automated red teaming be considered, to include AI-enabled capabilities. For example, CALDERA™[20] aids cybersecurity practitioners to automate cybersecurity assessments, leveling the asymmetric playing field between hacker and defender. These capabilities would benefit from increased research and development funding.

12 Increase federal funding to create common vocabulary and frameworks for AI alignment, and use those to guide future research.

As described in earlier sections, we do not recommend limiting or regulating scientific advancements in AI that move us toward AGI, as restrictions will also limit positive advancements with the potential for great societal benefits and are likely impractical to implement globally. However, creating AI alignment in systems as scientific and engineering advances are made can mitigate the risk of either humans tasking AI to carry out dangerous actions or AI systems allowing themselves to have emergent behavior. AI alignment research is nascent, and can be accelerated with additional research funding, first focused on establishing a common vocabulary and framework to align the research community and to identify research needs. Such a vocabulary and framework could subsequently be used by program managers who guide federally funded AI research to evaluate proposals' compliance with alignment principles, similar to guidelines established for research involving human subjects.

13 Regulation and legal frameworks should differentiate between appropriate research with risk mitigations and bad actors, and hold all appropriately accountable for harms.

We recognize that purpose is an inherently human quality, and AI systems with agency will either directly get purpose from a human (as instruction) or infer purpose through learning from human behaviors and artifacts. This is an area of active research. Advancements in AI alignment thinking and practices may serve to limit emergent, undesirable AI behavior, but research activities will still need safe environments with regulated guidelines similar to bioresearch and biosafety levels. And bad actors will undoubtedly try to create autonomous AI systems with malintent (e.g., Chaos-GPT). Similar to the argument on page 10, policy and regulation can go further to hold accountable and impose penalties on those that either unintentionally (researchers) or intentionally (bad actors) give self-direct AI systems goals that lead to harm.

About MITRE

MITRE's mission-driven teams are dedicated to solving problems for a safer world. Through our public-private partnerships and federally funded R&D centers, we work across government and in partnership with industry to tackle challenges to the safety, stability, and well-being of our nation.

About the Authors

T. Charles Clancy, Ph.D., is the senior vice president and general manager, MITRE Labs, and chief futurist, the MITRE Corporation. He previously served as MITRE's vice president for intelligence programs and, before that, as the Bradley Distinguished Professor in Cybersecurity at Virginia Tech and executive director of the Hume Center for National Security and Technology.

Douglas P. Robbins is the vice president of engineering in MITRE Labs. He leads MITRE's Innovation Centers across a wide range of technologies, including electronics, communications, systems engineering, and artificial intelligence. Previously, he served as MITRE's vice president for Air Force programs. Prior to that assignment, he led the strategic development of MITRE's Massachusetts-based operations, including new partnerships with the local high-tech ecosystem.

Ozgur Eris, Ph.D., is the managing director of the Artificial Intelligence and Autonomy Innovation Center in MITRE Labs. He leads over 200 artificial intelligence engineers and scientists who are focused on catalyzing the consequential use of artificial intelligence for the public good. Previously, he served as the distinguished chief engineer of the AI and Autonomy Innovation Center and also established its AI-enhanced Discovery and Decisions department.

Lashon B. Booker, Ph.D., is a senior principal scientist in MITRE Labs' Artificial Intelligence and Autonomy Innovation Center. He has published numerous technical papers in the areas of machine learning, adaptive behavior, and probabilistic methods for uncertain inference. He has served on the editorial boards of several journals, and regularly serves on the program committees for conferences in these areas.

Katie Enos is a senior principal on MITRE's government relations team. Prior to joining MITRE, she spent several years consulting for C-suite executives following an almost two-decade career working on Capitol Hill, including over a decade as the chief of staff to a member of the United States Congress.

Acknowledgments
The authors would like to thank Duane Blackburn; Jason Duncan; Nitin S. Naik, Ph.D., Josh Harguess, Ph.D.; Christine Callsen; Christina Liaghati, Ph.D.; and EJ Hillman for their thoughtful input and review of this document.

Endnotes

[1] The MITRE Corporation, "MITRE ATLAS," The MITRE Corporation, April 2023. [Online] Available: https://atlas.mitre.org/. [Accessed 2023].

[2] The MITRE Corporation, Microsoft and MITRE Create Tool to Help Security Teams Prepare for Attacks on Machine Learning Systems, McLean, VA: The MITRE Corporation, 2023.

[3] E. Ozoani, M. Gerchick and M. Mitchell, "Model Cards," 20 December 2022. [Online]. Available: https://huggingface.co/blog/model-cards. [Accessed 2023].

[4] National Institute of Standards and Technology (NIST), "AI Risk Management Framework," NIST, 2021. [Online]. Available: https://www.nist.gov/itl/ai-risk-management-framework. [Accessed 2023].

[5] K. Zetter, "An Unprecedented Look at Stuxnet, the World's First Digital Weapon," Wired.com, 3 November 2014. [Online]. Available: https://www.wired.com/2014/11/countdown-to-zero-day-stuxnet/. [Accessed 2023].

[6] K. Wiggers, "What is Auto-GPT and Why Does it Matter?," Tech Crunch, 22 April 2023. [Online]. Available: https://techcrunch.com/2023/04/22/what-is-auto-gpt-and-why-does-it-matter/. [Accessed 2023].

[7] J. A. Lanz, "Meet Chaos-GPT: An AI Tool That Seeks to Destroy Humanity," Yahoo!, 13 April 2023. [Online]. Available: https://finance.yahoo.com/news/meet-chaos-gpt-ai-tool-163905518.html. [Accessed 2023].

[8] E. Yudkowsky, "Pausing AI Developments Isn't Enough. We Need to Shut it All Down," TIME Magazine, 29 March 2023. [Online]. Available: https://time.com/6266923/ai-eliezer-yudkowsky-open-letter-not-enough/. [Accessed 2023].

[9] Future of Life, "Policymaking in the Pause," 12 April 2023. [Online]. Available: https://futureoflife.org/wp-content/uploads/2023/04/FLI_Policymaking_In_The_Pause.pdf. [Accessed 2023].

[10] J. Togelius and G. Yannakakis, "Choose Your Weapon: Survival Strategies for Depressed AI Academics," Cornell University, 31 March 2023. [Online]. Available: https://arxiv.org/pdf/2304.06035.pdf [Accessed 2023].

[11] D. McCabe, "White House Pushes Tech CEOs to Limit Risks of AI," The New York Times, 4 May 2023. [Online]. Available: https://www.nytimes.com/2023/05/04/technology/us-ai-research-regulation.html. [Accessed 2023].

[12] M. Baker, "Nuclear Arms Control Verification and Lessons for AI Treaties," Cornell University, 8 April 2023. [Online]. Available: https://arxiv.org/pdf/2304.04123.pdf. [Accessed 2023].

[13] S. Nellis and J. Lee, U.S. officials order Nvidia to halt sales of top AI chips to China, San Francisco: Reuters, 2022.

[14] E. Yudkowsky, "Pausing AI Developments Isn't Enough. We Need to Shut it All Down," TIME Magazine, 29 March 2023. [Online]. Available: https://time.com/6266923/ai-eliezer-yudkowsky-open-letter-not-enough/. [Accessed 2023].

[15] E. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang and W. Chen, "LoRA: Low-Rank Adaptation of Large Lan-Guage Models," Cornell University, 16 October 2021. [Online]. Available: https://arxiv.org/pdf/2106.09685.pdf. [Accessed 2023].

[16] A. Folberth, J. Jahnel, J. Bareis, C. Orwat and C. Wadephul, "Tackling Problems, Harvesting Benefits—A Systematic Review of the Regulatory Debate Around AI," Cornell University, 7 September 2022. [Online]. Available: https://arxiv.org/abs/2209.05468. [Accessed 2023]

[17] N. Leveson, "System Safety Engineering: Back to the Future," Massachusetts Institute of Technology Aeronautics and Astronautics Department, 2002. [Online]. Available: http://sunnyday.mit.edu/book2.pdf. [Accessed 2023].

[18] N. Leveson, "The Use of Safety Cases in Certification and Regulation," Massachusetts Institute of Technology Engineering Systems Division, November 2011. [Online]. Available: https://dspace.mit.edu/handle/1721.1/102833. [Accessed 2023].

[19] J. Rushby, "The Interpretation and Evaluation of Assurance," SRI International, Menlo Park, 2015.

[20] The MITRE Corporation, "CALDERA," The MITRE Corporation, 1997-2023. [Online]. Available: https://caldera.mitre.org/. [Accessed 2023].

MITRE's mission-driven teams are dedicated to solving problems for a safer world. Through our public-private partnerships and federally funded R&D centers, we work across government and in partnership with industry to tackle challenges to the safety, stability, and well-being of our nation.

MITRE | SOLVING PROBLEMS FOR A SAFER WORLD