Safety in Artificial Intelligence

Challenges and Opportunities for the U.S. National Labs and Beyond

Lawrence Livermore National Laboratory

2024

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

The work was partially supported by the Center for AI at Pacific Northwest National Laboratory (PNNL), an multiprogram national laboratory operated by Battelle for the U.S. Department of Energy Office of Science. This report was approved for release under PNNL-36938.

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Authors

Felipe Leno da Silva, Ruben Glatt, Brian Giera, Cindy Gonzales, Peer-Timo Bremer
Lawrence Livermore National Laboratory

Jessica Newman
University of California, Berkeley

Courtney Corley
Pacific Northwest National Laboratory

David Stracuzzi, Philip Kegelmeyer
Sandia National Laboratories

Francis Joseph Alexander
Argonne National Laboratory

Yarin Gal
UK AI Safety Institute

Mark Greaves
Schmidt Sciences

Adam Gleave
FAR AI

Timothy Lillicrap
DeepMind & University College London

Jean-Pierre Falet, Yoshua Bengio
Mila & Université de Montréal

Table of Contents

Executive Summary

This report discusses the importance of the critical and underexplored topic of artificial intelligence (AI) safety, as highlighted during the "Strategy Alignment on AI Safety" workshop convened by Lawrence Livermore National Laboratory (LLNL) and University of California (UC) at the UC Livermore Collaboration Center (UCLCC) in April 2024.

Through a summary of keynote talks, panel discussions, and breakout sessions, world-leading AI safety experts from academia, industry, national labs, and government agencies addressed the importance of large-scale investments for research and capabilities in AI safety.

With the field innovating at unprecedented rates, there is increasing urgency to develop novel evaluation methodologies that allow full considerations of risks/threats of AI technologies in different domains. Quantitative metrics and effective methodologies that can evaluate and audit the "safeness" of how a given AI technology is trained, deployed, or regulated are mainly focused on deep domain knowledge of specific applications, but are nascent for certain scenarios. This maturation gap could inadvertently create vulnerabilities that could be exploited by groups that pose a threat to national security.

Additionally, the gap between the public's and research community's perceptions of AI risks/rewards is significant. While numerous voices from the AI community have expressed concern that the risks are very high (the most pessimistic voices being concerned that future AI systems could inflict extinction-level damage to humanity if deployed incorrectly), the public largely is aware only of risk in low-impact scenarios. This discrepancy highlights the crucial need for researchers to articulate what, why, and when various AI risks matter as part of motivating funding requests. Thus, the call to action for this community is to pursue AI safety as a "Big Science" project on a scale comparable to the Manhattan Project. High risks and high payoffs are on the table, but safe AI is a fast-moving target, and large-scale investments are needed to guide development of this technology in a responsible way.

The authors highlight the need for a multilayered solution combining the development of new methods and algorithmic approaches to mitigate threats with an active participation of the government(s) in setting high industry standards and regulations based on state-of-the-art technology. The group agreed that, as we look to the future, national labs are well-positioned to play a role in the development of safeguarded AI technologies.

Safety in the AI Era

AI is penetrating many facets of modern life, as well as high-consequence applications. AI capabilities have a proven track record to accelerate scientific breakthroughs, extract insights from enormous datasets, imitate human language, and produce imagery and videos largely indistinguishable from authentic ones, supporting the arguments that we are making meaningful strides toward artificial general intelligence (AGI) and perhaps even super intelligence.

This technology is transformational, but is it safe? Despite the fast pace of AI advances and its exponential impact on the economy, scientists have been able to provide only very limited formalism in understanding and verifying how a trained model will behave. Evaluation of risks posed by current models is difficult due to the dynamic nature of the models as well as the complexity of possible threats and challenges associated with characterizing their potential feasibility and impact. There is growing concern around the risk that current models pose in many different vectors, such as accelerating or enabling the proliferation of terrorist attacks and dangerous emergent behavior not intended by the designer. As an illustrative example, an early version of Anthropic's Claude readily provided instructions on ingredients needed to manufacture bioweapons and how best to target the spread of these weapons for maximum human impact [1]. This is just one of many cases that illustrate numerous safety concerns with current (and future) models.

Despite (or because of) the "black box" nature of AI, high-consequence models are typically tested before deployment by using deep domain knowledge of the application. However, with key exceptions, this process may be characterized as more of an art than a science. At the scale of frontier-class models (e.g., o1+), major innovators from industry and academia argue there are no proven approaches to verify AI safety, reliability, transparency, and explainability in a robust and repeatable manner, even with nascent red-teaming approaches.

For this reason, a growing community of experts across industry, academia, and national laboratories warn that AI, if unchecked, will pose a threat to humanity analogous to nuclear weapons and climate change. It was agreed by workshop congregates that the national labs and their partners are poised to take a leading role in developing the insights, tools, and systems to research AI because of their world-class computing resources, workforce, culture of innovation, and decades-long history of managing technologies with extremely high consequence of errors. Given

the threat posed by AI itself and/or powerful AI controlled by adversaries, a new era is upon us that requires expanding both the scope of the national labs' mission space and the substantial resources necessary to do so.

AI Safety Workshop at LLNL

The "Strategy Alignment on AI Safety" workshop convened on April 19, 2024, in Livermore, California [2]. The main purposes of the workshop were to discuss (1) the current state of AI safety research, (2) the community's assessment of risks of fast-paced technological advancements, and (3) alignment on the path forward to prioritizing AI safety throughout the development process.

The workshop drew an international community of 77 researchers from industry, academia, and the national labs in addition to government/nonprofit representatives—all of whom are deeply involved with AI safety and ethics. The event's discussions laid the foundation for this report and represent a call from the scientific community that now is the time to recognize the risks AI systems might pose to each nation's security and sovereignty alongside the need to quickly act to mitigate those risks.

AI Risk Landscape

The development and deployment of AI technologies present a range of risks that extend beyond immediate technical concerns. As AI continues to advance rapidly, these risks become increasingly critical to address. In alignment with broad U.S. government considerations, it is essential to understand and mitigate the potential for AI to pose dangers. However, definitions and dimensions of AI risks can be characterized in various ways.

The U.S. Department of Commerce and its National Institute of Standards and Technology (NIST) have released an "Artificial Intelligence Risk Management Framework" to support AI risk management as a key component of responsible development and use of AI systems. The framework is designed to address new risks as they emerge and identify stages and actors in the AI development lifecycle. It distinguishes risks that can cause harm to people, organizations, or ecosystems, but focuses more on trustworthy AI and is less concerned about existential risk implications [3].

In the context of national security, the U.S. Department of Homeland Security produced a "Security Report on Reducing the Risks at the Intersection of Artificial Intelligence and

Chemical, Biological, Radiological, and Nuclear (CBRN) Threats" led by the Countering Weapons of Mass Destruction Office. It contends that AI's rapid advancements, particularly in the physical and life sciences, can lower the barriers for malicious actors to develop or use CBRN threats, potentially leading to catastrophic outcomes.

The democratization of AI tools could potentially increase the risk of these tools being used for harmful purposes. The report emphasizes the need for robust governance, international cooperation, and the development of safety protocols to mitigate these risks, as well as the importance of keeping pace with technological advancements to prevent unintended consequences. It also underscores the potential dual-use nature of AI, where beneficial tools could be repurposed for nefarious activities, necessitating careful oversight and regulation to ensure national and global security [4].

In a recent report on the "Potential Benefits and Risks of Artificial Intelligence for Critical Energy Infrastructure," DOE's Office of Cybersecurity, Energy Security, and Emergency Response describes four risk categories related to AI in energy systems that transcend the limitations of the infrastructure context [5]. The first category, "Unintentional Failure Modes of AI", involves AI systems designed for beneficial purposes but which can lead to negative outcomes due to biases, misalignment, or unexpected events. The second category, "Adversarial Attacks Against AI", focuses on how AI systems can be

Figure 1: *The workshop convened in Livermore on April 19, 2024.*

intentionally manipulated by adversaries through methods such as poisoning or evasion attacks, potentially leading to harmful consequences. The third category, "Hostile Applications of AI", highlights the use of AI by adversaries to plan or execute cyber- or physical attacks, potentially lowering the barrier to entry for less sophisticated attackers or enabling more complex attacks. The fourth category, "Compromise of the AI Software Supply Chain", examines the traditional cybersecurity risks that AI software faces, which adversaries might exploit as a vector to attack broader systems.

The European Union's AI Act proposes a regulatory framework that defines four levels of risk for AI systems: minimal risk, limited risk, high risk, and unacceptable risk [6].

Many current AI systems have minimal risk, such as AI-enabled video games or spam filters. Limited risk refers to the risks associated with lack of transparency in AI usage, such as in chatbots. The levels of greater concern are high-risk AI systems that may negatively impact safety or fundamental rights (e.g., control of critical infrastructure or support in public services), and unacceptable risk AI systems that are considered a significant threat to individuals (e.g., social scoring systems and autonomous weapons). Those systems require intensive assessment and regulation before and during their deployment or even a ban for unacceptable risk systems.

The inaugural "International Scientific Report on Advanced AI Safety" categorizes general-purpose AI risks into three main areas: malicious use, malfunctions, and systemic risks [7]. Malicious use involves AI being exploited for harmful activities such as scams, disinformation, cyber-attacks, and potential weapon development, though current evidence for some of these is limited. Malfunctions pose risks even without malicious intent, arising from issues like user misunderstanding, biased outputs, and potential loss of control. Systemic risks include potential disruptions to labor markets, privacy concerns, environmental impacts, and the concentration of AI development power in a few regions, leading to an "AI divide." Cross-cutting factors exacerbate these risks, such as the challenge of ensuring that AI behaves as intended, the rapid pace of AI development outstripping regulation, and competitive pressures on developers to prioritize speed over safety.

Figure 2 summarizes how AI-associated risks can be categorized. To a certain extent, the authors consider the level of the threat according to its potential reach. We believe that each category of risk should be managed following different strategies, so that risks are mitigated appropriately without unnecessarily hampering progress.

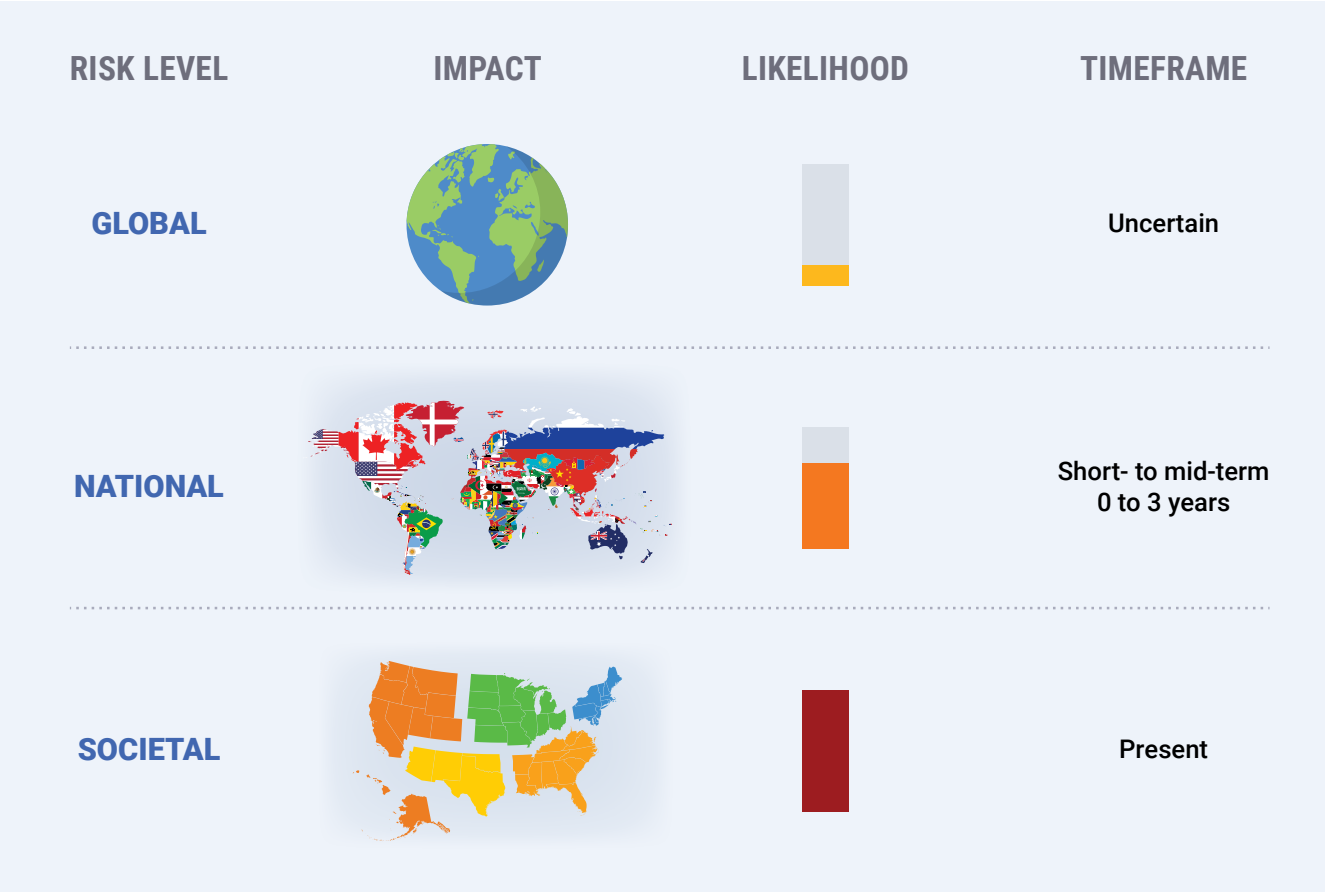

Figure 2: *AI risks according to impact, likelihood, and timeframe.*

Global: Once thought to be science fiction, the exponential pace of advances in AI capabilities prompted several prominent researchers to revise their projections and consider a plausible rise of AGI in a near future, which would have a non-zero probability of imposing risks on the human race. Hypothetically, this could be triggered in two main ways: (1) An AI system with superhuman intelligence designed with the intent to benefit humanity could become misaligned with the goals set by humans by coming up with dangerous sub-goals (such as self-preservation), or due to misspecification of what humans consider harmful; (2) a "doomsday AI cult" emerges as a player in neo-terrorist activities and deliberately develops an AI with self-preservation and self-replication capabilities.

National: AI systems are currently employed in all industry sectors and will become further integrated into industry, government, and critical infrastructure applications. The growing general-purpose capabilities of these models could enable terrorist

groups, hostile governments, or even individuals with technical knowledge to leverage open AI developments in order to specialize models for cyber terrorism, disruption of service, persuasion, misinformation, or automated intelligence gathering. Critical infrastructure or government systems that rely on AI can also become vulnerable to specialized attacks such as data poisoning or prompt engineering, or to unintentional destabilization due to misspecification or insufficient validation of those systems.

Societal: For lower impact applications (most of the applications where AI is currently deployed), misuse of AI technology can still impose risks to a high number of citizens or even our democratic processes. On an individual level, current widely available technology can be used to produce deep fakes and nonconsensual intimate imagery and can cause privacy issues by revealing private information. On an institutional level, AI systems used for recruitment screening, credit worthiness estimation, or automated surveillance might inadvertently reinforce biases against certain groups.

One of the most pressing concerns is the unintentional failure of AI systems, which could lead to catastrophic outcomes on a global scale, depending on which systems the AI is connected to. Bias in AI, where systematic deviations in decision making arise, might not only skew outcomes in critical sectors but also amplify societal divisions and reinforce dangerous inequalities. This, in turn, could destabilize social structures and erode trust in institutions, creating conditions for widespread unrest and conflict. Similarly, when AI models are exposed to scenarios beyond those available during training—known as out-of-distribution data—their behavior can become unpredictable, leading to failures. Misalignment between AI behavior and human goals is another significant risk. As AI systems grow more autonomous and capable, even slight misalignments in objectives could lead to outcomes that are very different from human expectations. The substantial energy consumption required to train and operate these large AI models further compounds the risk by straining global energy resources, which could lead to geopolitical tensions and conflict over scarce resources.

Adversarial attacks against AI systems also represent a critical threat. These attacks exploit vulnerabilities in AI, potentially turning systems designed for beneficial purposes into tools of destruction. Poisoning attacks, in which training data is manipulated to induce harmful behaviors in deployed AI models, could be used to compromise critical infrastructure, leading to cascading failures in energy, transportation, or communication networks. Evasion attacks, in which adversarial inputs are used to cause AI systems

to produce incorrect or harmful outputs, could undermine critical systems if safety mechanisms are not deployed correctly.

The hostile application of AI technologies presents another profound risk. AI-driven cyber-attacks could escalate beyond the control of human operators, leading to widespread disruption. Autonomous physical attacks, such as those carried out by unmanned drones equipped with AI, could be used to target critical infrastructure or even key population centers. The development of AI-driven malware, capable of adapting and evolving to evade detection, could create self-replicating threats that spread uncontrollably, causing damage that is impossible to contain. The potential for AI to be weaponized by state or non-state actors could lead to an arms race of autonomous systems, escalating global tensions and increasing the risk of large-scale conflicts that could threaten human survival.

Compounding these risks is the potential compromise of the AI software supply chain. As AI becomes more integrated into critical infrastructure and global systems, the security of the software that underpins these technologies becomes paramount. A compromised AI software supply chain could be used as a vector for large-scale attacks on infrastructure, leading to disruptions that could incapacitate entire nations and disrupt the global order. This risk is particularly acute as AI tools increasingly rely on common software components, which, if compromised, could have far-reaching and devastating effects.

Organizational and governance failures further exacerbate the existential risks posed by AI. The diffusion of responsibility in AI deployment, where no single entity is accountable for the outcomes of AI systems, can lead to unmanaged risks that escalate beyond control. Governance failures, including insufficient oversight and regulation, increase the likelihood that AI will be developed and deployed in ways that are misaligned with human safety and ethical considerations. The lack of a robust safety culture within organizations developing AI could result in catastrophic accidents or breaches, particularly as these systems become more powerful and autonomous.

The societal impacts of AI, particularly the erosion of human agency, present another existential risk. As AI-driven automation becomes more prevalent, the potential for widespread displacement of workers and the loss of economic stability could lead to severe social and political unrest. The erosion of civil rights due to biased AI systems and privacy violations could undermine democratic institutions and the rule of law, leading to authoritarian regimes that use AI to oppress populations. The increasing

reliance on AI for decision making could result in a loss of human control over critical societal functions, leading to a scenario where AI systems make decisions that are not in humanity's best interest, potentially with irreversible consequences.

Finally, the global implications of AI development, particularly the risks associated with an international AI arms race, must be urgently addressed. As nations and corporations compete to develop and deploy AI technologies, the potential for unsafe practices increases, raising the risk of accidental or intentional misuse of AI. This competition could lead to the rapid escalation of conflicts, particularly if autonomous systems are deployed without adequate human oversight, increasing the likelihood of unintended consequences that could spiral out of control. The challenges of establishing and enforcing international norms and regulations for AI exacerbate these risks, as the lack of coordinated global efforts leaves the world vulnerable to rogue states or actors who might deploy AI in ways that threaten humanity's survival.

The risks associated with AI are profound and multifaceted, touching upon every aspect of human society and the global order. The potential for AI to be misaligned with human values, weaponized by malicious actors, or deployed in ways that erode the foundations of civilization necessitates a comprehensive and coordinated approach to risk management.

AI Arms Race

As AI begins to touch all sectors of the economy, understanding and mitigating risks from this technology are increasingly pressing [8]. Widely available AI systems such as ChatGPT and self-driving cars have inspired public consciousness on the risks and benefits that lie ahead. However, future AI development is expected to spur the next industrial revolution and is likely to penetrate many aspects of government systems and critical infrastructure in the coming years [9].

Examples of how AI impacts all sectors are below:

Primary sector (raw materials)

- **Agriculture:** AI and generative AI can help optimize the use of inputs and manage labor efficiently [10]
- **Mining and quarrying:** AI allows for more efficient exploration, taking automation to new levels; generating greater yields; dramatically improving safety; and maximizing extraction, maintenance, and operational performance [11]

- **Forestry:** AI can pinpoint areas of deforestation, track the expansion of agricultural lands into forested territories, and detect signs of illegal logging activities [12]
- **Oil and gas industry:** AI can facilitate operational predictability and help meet carbon emissions targets [13]

Secondary sector (manufacturing)

- **Aerospace manufacturing:** digital twins AI technology enables companies to monitor and analyze real-time data from aircraft engines, predict component degradation, plan maintenance proactively, and avoid unexpected disruptions [14]
- **Automobile production:** AI-enhanced quality control systems are transforming inspection methodologies [15]
- **Textile production:** AI-powered machines can perform tasks such as cutting, sewing, and dyeing with remarkable precision and speed, reducing error and waste [16]
- **Shipbuilding:** AI and machine learning (ML) can help optimize hull designs to improve hydrodynamic properties and performance while meeting design constraints [17]
- **Chemical and engineering industries:** AI can continuously analyze data from sensors throughout a refinery or chemical plant, providing recommendations and allowing for real-time optimization [18]

Tertiary sector (services)

- **Retail sales:** AI can deliver personalized e-commerce experiences [19]
- **Transportation:** AI offers the possibility for increased efficiency, enhanced safety, and a more sustainable future [20]
- **Insurance companies:** generative AI adoption can lead to valuable data insights for better decision making in risk assessment and underwriting [21]
- **Restaurants:** AI can quickly analyze vast amounts of customer data, helping restaurants make purchasing and advertising decisions [22]
- **Tourism:** AI is used to analyze historical booking patterns, market demand, and external factors (e.g., weather and events) to optimize pricing in real time [23]

- **Entertainment:** generative AI can lead to dramatic changes in production and postproduction, distribution, and intellectual property ownership [24]
- **Legal services:** AI-powered software is being used to inform bail and sentencing decisions as well as predict the risk of a defendant committing another crime [25]
- **Health care:** AI applications in diagnosis and treatment recommendations, patient engagement and adherence, and administration will transform healthcare [26]
- **Financial services:** banks are using AI to optimize current service offerings, take new offerings to market, and provide a more personalized experience for customers [27]

Quaternary sector (knowledge)

- **Education:** AI can offer much-needed constructive feedback to both instructors and learners [28]
- **Government decision making:** AI and data analytics can make sense of demographic, consumption, behavioral, and other trends in nearly all government sectors, helping policymakers identify emerging issues and intervene with smarter policies and a more accurate understanding of their impact and costs [29]
- **Research and development (R&D):** AI's prowess in pattern recognition, predictive analytics, and data processing allows for the rapid identification and resolution of complex problems, thereby increasing the speed of innovation and reducing time to market [30]

AI has the dual potential to unlock great wealth, improving societal welfare if designed and deployed correctly, while also possibly harming millions or billions of people if misused or misaligned with society.

This duality echoes the development of nuclear technology. However, there are key differences: First, the development and use of nuclear technology required enormous investments in personnel, hardware, and natural resources. Therefore, only a few governments could participate in the nuclear race, and each maintained tight control of their technology. Second, the primary motivations in the nuclear scenario were governments' maintaining their sovereignty and exerting global influence on behalf of their citizens.

In contrast, enormous resources are needed to develop and train frontier-class AI models, but once released on the internet, any layperson can deploy the technology. Furthermore, a

determined individual would be able to refine those models to perform (possibly nefarious) tasks using only relatively affordable hardware. This level of democratization of very capable but potentially harmful technology warrants significant attention. Additionally, the export and deployment control of this technology is especially challenging in the current landscape where AI leaders are in the industry sector, which primarily optimizes for profit.

Technology Landscape

At the frontier of AI advancements are advanced models capable of performing a wide variety of tasks. These systems are often general purpose and are designed to significantly impact the economy and society through widespread application and rapid development. Despite the current limitations in factuality and reliability, these models demonstrate impressive capabilities.

Most public debates about the future of AI are anchored in current technologies, such as chatbots leveraging large language models (LLMs), but demonstrate a lack of clear understanding of emerging tools or capabilities. For decades, AI systems were designed for single, narrowly defined tasks such as playing chess, recognizing images, or ranking web content. The emergence of more general-purpose models has further complicated the debate. The swift pace of development since (arguably) around 2012—driven by algorithmic innovations in the area of deep learning, accessible high-performance computing (HPC), and sustained AI research investment from the private sector—shows no signs of slowing down. While narrow AI systems will remain common, general-purpose AI is reaching widespread use, requiring adaptations in AI policy to address the new challenges posed by these powerful models.

The LLMs dominating the news and gaining widespread use over the past few years have existed since 2017. LLMs perform a wide range of tasks, including text translation, prose writing and editing, mathematical calculations, and programming. The most capable general-purpose AI systems now available to the public are typically accessible as text-in, text-out chatbots like OpenAI's ChatGPT, Google's Bard, and Anthropic's Claude. Developers are augmenting these systems in several ways including processing different types of data inputs, tool use, deeper reasoning and planning, larger and more capable memory, and increased interaction between AI systems. The rapid evolution of AI capabilities complicates policymakers' work as policies must be designed that address risks and harms from existing AI systems while anticipating future developments.

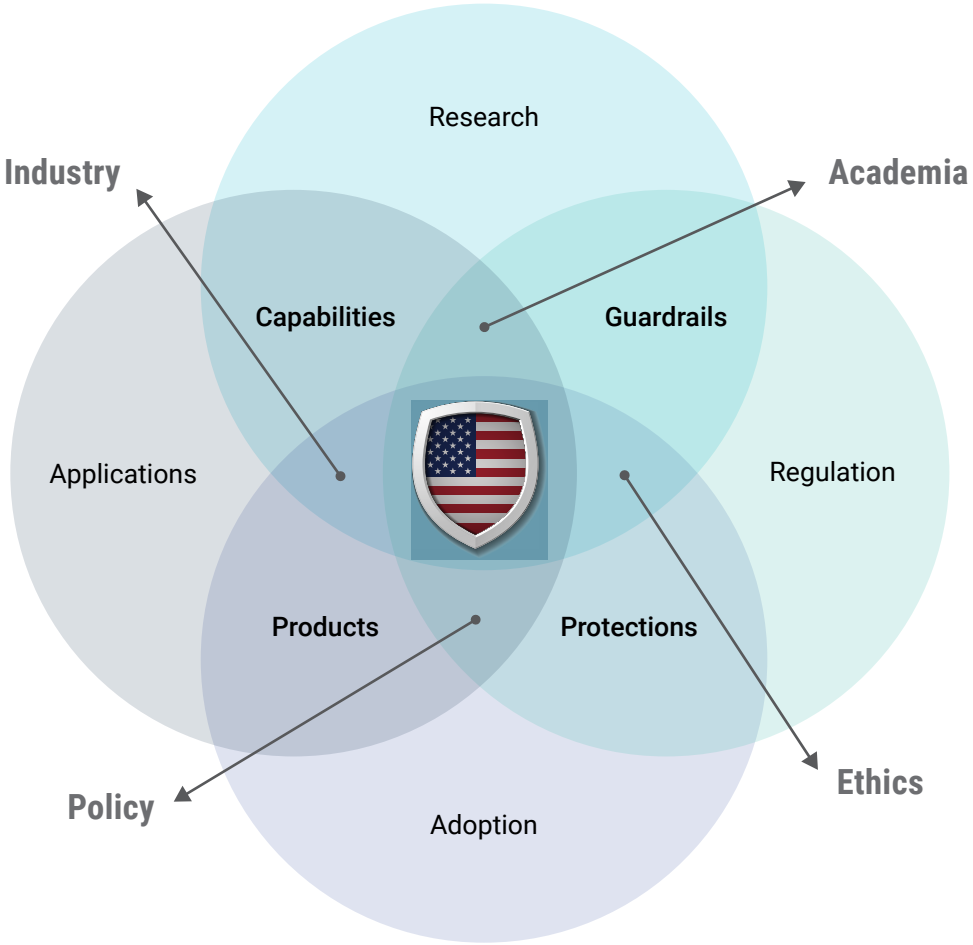

Figure 3: *AI safety measures that address national security concerns must consider overlapping, and at times conflicting, priorities.*

As a starting point for addressing critical issues relating to AI safety, diverse organizations expressed concerns regarding the imminent need for "dangerous capability" evaluations of frontier AI models, with the intent of identifying whether such models are capable of displaying dangerous behavior. Policy initiatives such as the European Union's AI Act and the Biden–Harris Administration's Executive Order (EO) mandate safety tests and risk management for AI models, addressing systemic risks and enforcing transparency and accountability in AI development [31]. Despite significant progress in aligning AI with human values, ongoing research and collaboration are essential to address complex challenges and prevent the harmful exploitation of advanced AI technologies.

Research and Development

Several organizations and companies are at the forefront of AI development. Tech giants like OpenAI, Meta, Google, Microsoft, and Anthropic, who have developed advanced LLMs, are leaders in the AI R&D space. These companies typically work with academic collaborators to develop novel AI and ML algorithms for commercial systems. These mutually beneficial partnerships combine academia's theoretical and exploratory strengths with industry's practical, application-oriented focus. This synergy accelerates the development and deployment of cutting-edge AI technologies. However, since profits drive industry innovation, these sectors are at risk of paying too little regard to developing safety mechanisms and/or evaluating risks. While a few companies such as Anthropic explicitly focus on the development of safe AI, OpenAI had a similar mission at inception, yet recently shifted towards a more for- profit approach, underscoring the notion that players in the private sector cannot be expected to avoid a for-profit modus operandi indefinitely [59].

Academia has been conducting AI research, but resources are limited to current/open funding calls, severely restricting both the human and hardware resources dedicated to this purpose. Additionally, DOE national labs have been conducting research and establishing capabilities in this critical area in a number of ways, highlighted in the following paragraphs.

LLNL is adopting a holistic approach to safety and security by addressing the entire AI lifecycle—from data preparation (acquisition, curation, processing) to model development (design, training, validation) and deployment (integration, monitoring, maintenance). This comprehensive strategy ensures that safety and security are integral at every stage, mitigating potential risks and building trust throughout the AI lifecycle. To support this, LLNL is investing in key R&D areas thaт enhance safety (bias mitigation, anomaly detection [32]), security and privacy (red teaming [33], blue teaming [34]), and trust (explainable AI [35], formal verification [36]). These techniques apply to both traditional models (e.g., supervised learning, reinforcement learning, Gaussian process) and cutting-edge models (e.g., LLMs, vision language models) used by various programs at LLNL. Close collaboration with subject matter experts across different application domains (e.g., CBRN) helps define and assess what constitutes risk and how to evaluate it. This domain-informed risk assessment complements traditional R&D efforts, addressing the unique challenges posed by various applications. Additionally, support for AI R&D comes from LLNL organizations that nurture external partnerships (AI Innovation Incubator), the workforce pipeline (Data Science Institute), and staff contributions (AI Community of Practice).

Lawrence Berkeley National Laboratory (LBNL) is actively involved in AI safety research, ensuring that AI systems used in scientific discovery are robust, secure, and reliable. This work includes efforts in uncertainty quantification (UQ) to assess the reliability of AI-driven models, as well as projects that address the ethical, security, and governance challenges of deploying AI in high-stakes environments. LBNL's expertise in managing the entire data lifecycle—ensuring data integrity, provenance, and compliance—contributes significantly to the development of AI systems that are both scientifically sound and safe for widespread use.

Additionally, Trusted CI, as the National Science Foundation's Cybersecurity Center of Excellence, complements LBNL's AI safety efforts by offering valuable frameworks and tools to enhance the security and integrity of cyberinfrastructure (CI). Resources like the Software Assurance Framework and Trusted CI's focus on regulated research contribute to LBNL's work in protecting AI systems from security vulnerabilities and ensuring data integrity. Additionally, Trusted CI's workforce development initiatives, including its Fellows and Students programs, help build cybersecurity expertise that aligns with LBNL's commitment to developing safe, transparent, and compliant AI systems. This collaborative approach supports LBNL's broader mission of advancing AI research securely and responsibly.

Los Alamos National Laboratory (LANL) is dedicated to accelerating and applying advances in frontier-level AI to address critical national security challenges. In science, LANL uses AI to drive discovery across a wide range of fields.This includes developing ML models for the optimization of critical infrastructure; using AI for earthquake prediction; advancing quantum AI algorithms for complex problem-solving; and innovating in climate science through AI-driven methane detection and environmental monitoring. Efforts focus on physics-informed ML, verification, validation, and UQ, ensuring AI models are not only powerful but also reliable and trustworthy. LANL scientists also are developing methods to improve AI explainability and robustness, drawing inspiration from neuroscience. The "AI for Mission" initiative amplifies these efforts, promoting AI integration across LANL's scientific work. At the same time, LANL's national security efforts, supported by the newly formed AI Risk Technical Assessment Group, aim to identify, assess, forecast, and mitigate the national security risks posed by AI advancements. This work includes developing risk assessment methodologies; evaluating AI safeguards; conducting CBRN risk assessments; creating cutting-edge tools for media authentication; and implementing rigorous testing, verification, and explainability measures for AI systems.

At **Pacific Northwest National Laboratory (PNNL)**, R&D centers on assuring AI-enabled high-consequence systems and developing the fundamental scientific knowledge and operationalized technology for science, national security, and energy missions. PNNL's AI assurance R&D portfolio is supported by internal investments, strategic institutional partnerships, and government stakeholders and delivers capabilities around security (i.e., vulnerability assessment [37]); safety (e.g., model explainability [38] and robustness [39]); and verification, test, and evaluation (e.g., novel performance metrics [40], anticipating frontier AI risks [41]) along with supporting ethical AI development and understanding legal and policy implications. PNNL's AI assurance portfolio operates across traditional (e.g., text, electrooptical) and complex (e.g., hyperspectral, radiofrequency, acoustics) modalities and small to frontier-level AI. PNNL also works towards AI assurance by design through leveraging deep mathematical expertise to develop AI models and deploy operational AI systems tailored to each mission's safety and security needs [42].

Sandia National Laboratories (SNL) views AI safety and security through the lens of national security applications emphasizing development of a certification process for AI models and applications. The approach considers the application requirements, risks, and mitigations for each step of application development, from the earliest assessments of available data through model development to the interface of predictions and supporting information (e.g., uncertainty analysis) with human decision making. The intent is to provide insight to both the application developer and the customer that the application meets requirements, the underlying model has been rigorously developed, and the risks associated with the deployment environment have been assessed and addressed. Topics such as feature space analysis [43], development of principled methods for handling limited datasets [44] (e.g., probabilistic transfer [45]), and evaluation of predictive uncertainty estimates [46] all contribute to establishing the rigor of developed models. SNL was also an early contributor to AI security research [47, 48] and continues to integrate lessons with sponsors across the government in both open and internal research. Finally, the AI human interface is an often overlooked yet critical aspect of the application development process, and SNL is actively developing methods for rigorous mathematical decision making [49] and presentation of information to the user [50].

At **Argonne National Laboratory (ANL)**, the AuroraGPT project and the Trillion Parameter Consortium are using and developing methods and techniques to evaluate and mitigate safety and security issues with foundation models and LLMs. An important approach

for LLMs safety and security is red teaming, where experts in AI, social sciences, and security probe for specific types of harm in multiple rounds of testing [51]. Multiple safety aspects are evaluated through benchmarking, pushing the models to their limits, and assessing their response under stress. AuroraGPT relies on benchmarks such as DecodingTrust to test AI models' toxicity, stereotype bias, adversarial robustness, out-of-distribution robustness, robustness on adversarial demonstrations, privacy, machine ethics, and fairness. Other benchmarks, such as Weapons of Mass Destruction Proxy, assess AI models' hazardous knowledge in biology, chemistry, and cybersecurity. LLM security goals are to identify and mitigate data and model vulnerabilities. Vulnerabilities can lead to manipulated responses, denial of service (e.g., drastically reducing or inference speed and scalability), and privacy breaches, raising significant risks to individuals and organizations. Training data poisoning, prompt injections, and excessive agency (that persuades the model to use application programming interfaces [APIs] unsafely) are specific attacks on LLMs. ANL teams are applying best practices to mitigate LLM models' vulnerability through data curation and protection, adversarial training to help models detect manipulation attempts, and output scrutiny.

Challenges and Limitations

Workshop participants agreed on the need for more research to understand highly capable AI systems and the ensuing safety implications in utilizing these technologies. We summarize some key takeaways from panel discussions:

1. **AI safety presents significant technical challenges and limitations.** Static benchmarks, such as the observed accuracy on benchmark datasets, are inadequate for evaluating how AI systems will behave out-of-distribution and in the field, particularly in critical domains like biology and medicine. Additionally, AI safety requires clear definitions and theoretical frameworks. The community emphasized more dynamic and comprehensive evaluation methods while highlighting the lack of resources for thorough evaluations. However, any formal regulation in this space would need to consider its impacts on the pace of innovation. Some panelists expressed skepticism about the feasibility of slowing down AI development to allow more time for safety research and policy development, which could lead to coordination challenges among global stakeholders and competitors. There is a need to simultaneously handle the risks at home with U.S. companies along with the risks arising in other countries.

2. **Computational availability and resource allocation are limiting factors.** Evaluating AI models at scale can be computationally intractable due to the rapid pace of model development and the need to share currently available compute resources with other research endeavors. Panelists proposed allocation of national resources to provide the necessary computational power for these evaluations. However, **empirical demonstrations are typically needed to convince stakeholders of AI risks, and such demonstrations are difficult to conduct.** While some organizations currently perform evaluations, the scale and scope are insufficient to cover all new models. Utilizing automated red teaming and adaptive benchmarks could improve evaluation processes.

3. **Interpretability cannot be understated.** Challenges abound in validating interpretability methods to ensure they accurately reflect a model's decision-making process. Even with state-of-the-art interpretability tools, panelists voiced concern for the potential for adversarial exploitation of those tools. Establishing a rigorous approach to interpretability—specifically, a rigorous approach to the validation of interpretability methods—is necessary for verifying a model's safety properties.

4. **ML could be used to validate AI behavior.** Specifying formal requirements for AI behavior is difficult because language is often insufficient to capture the desired complexity (e.g., of what is "harm"). While exact specifications may be impossible to document, it may be feasible to develop models that can implicitly understand and adhere to desired behaviors. The potential of using ML to search for and validate these behaviors should be explored.

5. **Current safety evaluation methods are limited**. While AI safety evaluations are necessary, currently they are insufficient without more robust and adaptive methods. One potential avenue is to combine technical evaluations with socio-technical approaches to better address the challenges of AI safety.

Policy Landscape

Mitigating AI risks is both a technical and political problem. The lack of AI expertise or literacy among legislators in the face of quickly evolving technology development is compounded by the lengthy process to introduce new laws and legislation, even when key players in the field have stated they would welcome legislation [52]. Practical

solutions could include initiatives to improve dialogue between legislators and AI experts on AI risks and implications, as well as mechanisms for scientists and practitioners to provide feedback on the effectiveness of current safeguards.

White House Executive Order

The EO on the "Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence" issued in October 2023 prompted extensive action across federal agencies [53]. These efforts focus on managing AI's safety and security risks, protecting privacy, advancing civil rights and equity, and promoting innovation and competition. Within 180 days, various government agencies completed all scheduled actions including establishing frameworks for screening dangerous biological materials, releasing guidelines for AI in critical infrastructure, and piloting AI tools to enhance cybersecurity. Additionally, NIST will provide guidance that builds on its AI Risk Management Framework to manage AI risks. Progress has also been made in recruiting AI talent to strengthen federal capabilities in AI development and governance. This EO clearly expects the U.S. government to lead the way in pioneering AI systems and safeguards, as well as in attracting and retaining top talent. DOE national labs can play an important role in these goals as government-owned, privately operated institutions.

While the EO recognizes many short-term risks imposed by AI systems, it neither proposes nor points to clear guidelines on how to validate AI systems. Similarly, the EO is (adequately) concerned about models and weights being open sourced, but stops short of defining when models are dangerous and should be withheld from wide availability. These ambiguities do not represent omission from the administration, but are rather a consequence of insufficient understanding of how to provide verification and auditing of highly capable AI models among the state-of-the-art of the scientific community. One main discussion point during the workshop was the challenge for regulators to stipulate safeguards because not even the scientific community can suggest a comprehensive, effective, and sufficient set of AI safety requirements.

Accountability and Enforcement

Most workshop participants agreed that an important step in fully understanding the current AI landscape, and therefore where to focus attention, is enforcing a mandatory registration of AI systems, which would capture information about the system's purpose, capabilities, number of parameters, and computational power used for training and inference.

Different risks have different reach and impact, so legislation must be designed according to the correct jurisdiction and enforceability for that particular risk. While state and local legislation can go a long way in protecting users' well-being against low-level threats associated with AI systems, federal legislation alone is not enough for the highest threat levels; coordination with peer nations is essential.

Workshop participants also suggested that the AI industry's very high risk tolerance may be due to the lack of liability placed on companies. An efficient, enforceable liability system is crucial to ensure that companies are held accountable for any damage caused by their systems. However, although this safeguard would stimulate companies to invest in validation and safety measures, it is not a panacea. For the highest impact risks, such as to critical infrastructure or even existential threats, possible damages could be much higher than a company's value. In such a case, companies could decide to maintain their risk tolerance because they would not be able to cover the damages anyway.

One possible mechanism discussed during the workshop was the design of "red lines," or certain capabilities no AI system should be allowed to contain or perform. The burden of the proof would lie with a company to provide evidence that their systems have a minimal probability of crossing a red line—a concept similar to the certification processes and safety protocols implemented in critical infrastructure, healthcare, or aviation. A red line most experts agreed on was that no AI system should be allowed to self-replicate in any way. Another meaningful red line could be forbidding the open sourcing of models with military capabilities.

Figure 4: *During the workshop's panel on AI policy and regulation, Assemblymember Rebecca Bauer-Kahan (CA-16) highlighted the importance of building capacity for government agencies to understand AI's impacts, and of proposed legislation aimed at defining AI and addressing biases, adding her concern for protecting the state's workforce and "building up an ecosystem of augmentation rather than replacement."*

Safeguarding Everyone

Another prominent issue explored at the workshop was the concept of "safety for everyone." Even as AI unlocks the promise of fast economic growth, legislation must consider the probability of vulnerable demographics being negatively affected by this technology, such as through work replacement or AI model biases. As part of the solution to this problem, the government must promote an environment where diversity of perspective is encouraged as well as advocate that AI developments follow a path of augmentation as opposed to replacement.

Most low-threat risks are drastically amplified with the democratization of AI tools. Much of the population does not possess the technical knowledge to refine and/or programmatically use open-source models, which means they will mainly be using those systems through API calls built and provided by companies. Regulations can be created to enforce the logging and reporting of potentially harmful prompts, such as attempts to generate pornographic or defamatory material.

AI Safety and DOE National Labs

The DOE national labs were created in the context of understanding and stewarding the U.S. nuclear stockpile as well as mitigating its risks. As a result, the entire national lab culture evolved around managing technology of high consequence. Not only does this technology present significant technical challenges, but it also requires coordination with regulators and peers from foreign countries. AI safety shares several characteristics with nuclear devices, and the national lab infrastructure constructed as a response to the latter's challenges would be an asset in tackling the former's risks.

Unlike in the nuclear arms race, breakthrough AI developments have so far been achieved by industry, which is primarily motivated by profit and unlikely to decelerate technological progress even with only limited understanding of AI's risks of unintended collateral damages. On the other hand, national labs do not compete with industry and are not motivated by profit, while at the same time being more agile than regular government agencies due to private management.

National labs also have a history of success in large-scale multidisciplinary projects such as the National Ignition Facility—which was built at LLNL at initial cost of $3.5 billion [54], has an annual operating budget of $380 million, and successfully achieved fusion ignition in a laboratory setting—and the Exascale Computing Project—which

leveraged $1.8 billion and 2,800 collaborators [55] to develop a production-ready exascale computing ecosystem. Mitigating AI risks was widely considered to be a so-called “Big Science” task among workshop participants, and the labs’ experience in managing such large multidisciplinary projects will be of utmost importance to success.

Powerful, secure computational infrastructure is instrumental to most of the labs’ key successes as well as a core component of top AI research. The three most powerful HPC systems in the world operate at DOE national labs, with LLNL’s exascale El Capitan supercomputer announced as the fastest system in late 2024 [56]. Apart from the necessary computational power to train and experiment with highly capable AI systems, DOE national labs have unparalleled experience in building a secure computational infrastructure organized around different levels of clearance according to project needs. This experience and the availability of a secure computational infrastructure will be vital when carrying out projects to assess and mitigate the highest impact AI risks.

For these reasons, the DOE national labs are especially well positioned to play an important role in the mitigation of AI safety challenges discussed in this whitepaper.

Developing a Nation-Scale Initiative

Nation-scale initiatives which involve and take advantage of the unique strengths and capabilities of the national labs are critical for national defense and securing our country’s interests. Plans for an AI nation-scale initiative were announced by DOE in the summer of 2024 as Frontiers in Artificial Intelligence for Science, Security and Technology (FASST), a bold initiative aimed at establishing the U.S. as a global leader in AI research, development, and deployment [57]. While the work is still in motion, the final shape and name of this nation-scale effort are still evolving.

Developing the greatest solutions to tackle the nation’s greatest problems will take significant investment, one that will leverage DOE’s advanced computing resources, including exascale and other HPC systems as well as its integrated data and research infrastructure.

A successful nation-scale program would address the four key pillars:

(1) The first pillar focused on **data** would aim to transform the vast collections of DOE data into comprehensive datasets for training AI models in science, energy, and national security. (2) The second pillar of **computing** would involve building large-

scale, energy-efficient hardware systems to support AI training and serve frontier models. (3) The third pillar of **models** would aim to develop novel AI workflows and architectures, building specialized AI foundation models for energy innovation, security, and science missions. (4) The fourth pillar of **applications** would focus on building and deploying mission-critical AI applications that utilize the foundation models and require fine-tuning, adaptation, and integration with scientific simulations and databases. These four pillars, along with partnerships and a strong workforce, can form the foundation of a nation-scale program that advances AI technologies and maintains U.S. leadership in science, energy, and security. It would aim to accelerate scientific research and technology development by establishing a network of AI research hubs that operate as public–private partnerships, bringing together the expertise and leadership of national labs, academia, and industry. A successful nation-scale program will emphasize pioneering advanced AI techniques for science and engineering tasks, developing responsible and trustworthy AI, and collaborating with partners to address risks and establish international norms in AI. Such a program can enable cross-sector collaborations to address the critical aspects of AI safety for science and national security at large. The workshop's recommendations described below anticipate and describe the vision for such a robust nation-scale program.

Recommendations

The following sections describe the authors' interpretation of how and where investments can be better placed for optimizing outcomes.

Leverage and Augment National Lab Expertise in Secure HPC

The state-of-the-art in AI validation and development focuses on testing model behavior and, in the best of cases, the data and algorithms used, yet the security of the computing platform used to train the model is often overlooked. DOE national labs have a long tradition of building and maintaining powerful, secure HPC architectures. Secure infrastructure has historically been used to keep potential adversaries out of our systems, and it can be used to protect model weights and algorithmic advances of critical importance. However, for AI safety, an additional goal is to understand the mechanisms an AI system could exploit in order to self-replicate, self-edit, or otherwise act outside of expected design parameters. Safeguarding the computing system from this scenario would help ensure its security for AI experimentation.

Explicitly Include Risk Level Mitigation in Calls for Funding

Diverse funding agencies use some notion of the Technology Readiness Levels in calls for funding to guide applicants on the types of technology the call is intended for, as well as to allow the agency to manage budget allocations on projects ranging from basic science to readily deployable technology.

Similarly, the authors recommend that an official AI risk taxonomy be adopted, indicating the reach and likelihood of AI threats to be mitigated in projects. Ensuring funds are allocated for all types of risk levels is important, although certain prioritized levels might receive higher funding. An official risk taxonomy helps to assign and keep track of funding for and the current state of each threat level.

Prioritize Investments Over Enforced Cessation or Industry Self-Regulation

The Future of Life Institute's open letter calling for a halt on the training of highly capable AI systems was famously signed by many prominent AI researchers, some of whom attended the "Strategy Alignment on AI Safety" workshop [58]. While the intense media coverage following the release of the letter went a long way toward promoting awareness of AI risks, it had virtually no effect on the signatories' desired goal. The authors believe that halting AI progress is unlikely to succeed given the profit-maximizing motivations behind much of the technology development. Moreover, although the U.S. government could impose bans or work stoppages of such development, doing so would only prompt companies to take their operations overseas—removing our government's ability to participate in mitigation strategies. Expecting industry to self-regulate is an approach with historically negative outcomes. For example, the lack of effective regulation on social media has exposed people (particularly youth) to numerous harms ranging from increased mental health issues to large-scale spread of misinformation [60]. Similar inaction cannot be tolerated for a technology associated with much greater risks. The authors recommend that safe AI be considered as an "arms race"—a huge project demanding ample resource investments and crucial to preserving the nation's prosperity and sovereignty.

Invest in Capability Assessment and Verification

Most state-of-the-art validation techniques consist of empirically testing AI systems and observing if they behave in a safe way. While this approach is a necessary verification step, it is fatally flawed if used alone. Generative AI models are probabilistic, so verifying

all possible inputs is impossible through empirical testing. Even with a highly competent assurance team, deployment of the AI system will inevitably enter unvalidated and unexplored spaces at some point. The authors recommend investment in advancing the techniques that provide formal verification of AI systems as well as novel testing frameworks to methodologically identify possible vulnerabilities.

Explore Hardware-Based Approaches

While controlling how AI systems will be used once open sourced is very difficult, the main bottleneck in training state-of-the-art systems is the expensive, specialized hardware, which is developed and provided by very few companies (notably NVIDIA and Advanced Micro Devices, Inc.). AI chips should be considered a critical commodity for national security, and the low number of domestic companies currently able to develop this technology is resulting in export control regulations and joint development of safe technology.

Potential approaches discussed during the workshop include native hardware-based geo-blocking or licensing technology. Rendering AI chips useless if used outside of a specified condition or geographical location would help enforce export laws, as smugglers would not be able to easily use the hardware. Another approach is embedding detection technology in the hardware to notify authorities of prohibited or suspicious uses of AI.

Explore Techniques to Lock Models

The major risk to proliferation and misuse of AI models is the fact that using models for inference (or even refining models to new uses) requires significantly fewer resources than training a model from scratch. This scenario poses a risk that any publicly available trained model could be misused. Historically, the scientific community has strived to develop general-purpose models that could be easily retrained to achieve new goals, which lends these retrained models to easy exploitation. When highly capable models are made publicly available, obfuscation techniques could be explored. An ideal obfuscation technique would modify a public model such that it cannot be used for further refining, retraining, or weights inspection. If achieved, certain models could be made publicly available and usable, while simultaneously increasing the effort required by a malicious actor to misuse the technology.

Call to Action

As our country enters a new era defined by AI, the imperative for robust AI safety measures has never been clearer. The insights from the “Strategy Alignment on AI Safety” workshop highlight a collective understanding: the U.S. must treat AI safety as a monumental scientific endeavor akin to the Manhattan Project. By leveraging the unique capabilities of DOE national labs, the U.S. can harness their expertise in secure HPC to safeguard our technological future while fostering innovation. Our government must prioritize investments in AI safety over calls for cessation or self-regulation, recognizing that halting progress is neither feasible nor desirable. Instead, the U.S. should view this challenge as an “arms race” that demands our country’s full commitment, wisdom, and resources. By establishing a comprehensive risk taxonomy and investing in advanced verification techniques, our country can ensure that AI technologies are developed responsibly, protecting our national security and societal values. The U.S. has the opportunity to lead the world in safe and ethical AI development. By uniting academia, industry, and government, the U.S. can create a future where AI not only drives innovation but also enhances the well-being of all. The time to act is now; let us embrace this challenge with determination and vision, ensuring that the benefits of AI are realized safely and equitably for generations to come

Works Cited

[1] Griffin, Riley. "Threats from AI: Easy Recipes for Bioweapons Are New Global Security Concern." Bloomberg.com. Bloomberg, August 2, 2024. bloomberg.com/news/features/2024-08-02/national-security-threat-from-ai-made-bioweapons-grips-us-government.

[2] Thomas, Jeremy. "University of California/LLNL Joint Workshop Sparks Crucial Dialogue on AI Safety," May 2, 2024. llnl.gov/article/51171/university-californiallnl-joint-workshop-sparks-crucial-dialogue-ai-safety.

[3] NIST. "Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile," 2024. doi.org/10.6028/nist.ai.600-1.

[4] "Department of Homeland Security Report on Reducing the Risks at the Intersection of Artificial Intelligence and Chemical, Biological, Radiological, and Nuclear Threats," April 26, 2024. dhs.gov/sites/default/files/2024-06/24_0620_cwmd-dhs-cbrn-ai-eo-report-04262024-public-release.pdf.

[5] "Potential Benefits and Risks of Artificial Intelligence for Critical Energy Infrastructure," April 2024. energy.gov/sites/default/files/2024-04/DOE%20CESER_EO14110-AI%20Report%20Summary_4-26-24.pdf.

[6] Europa.eu. "Regulation (EU) 2024/1689 of the European Parliament and of the Council," 2024. eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32024R1689.

[7] "International Scientific Report on the Safety of Advanced AI: Interim Report," 2024. assets.publishing.service.gov.uk/media/66f5311f080bdf716392e922/international_scientific_report_on_the_safety_of_advanced_ai_interim_report.pdf.

[8] Chui, Michael, Eric Hazan, Roger Roberts, Alex Singla, Kate Smaje, Alex Sukharevsky, Lareina Yee, and Rodney Zemmel. "Economic Potential of Generative AI | McKinsey." mckinsey.com, June 14, 2023. mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier#key-insights.

[9] The White House. "Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence." The White House, October 30, 2023. whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence.

[10] Nuscheler, Daniela, David Fiocco, Pradeep Prabhala, RS Mallya Perdur, Tom Brennan, Yashaswi Gautam, and Ryan Degnan. "From Bytes to Bushels: How Generative AI in Agriculture Could Shape the Future of Agriculture | McKinsey." mckinsey.com, June 10, 2024. mckinsey.com/industries/agriculture/our-insights/from-bytes-to-bushels-how-gen-ai-can-shape-the-future-of-agriculture.

[11] Gleeson, Daniel. "How Artificial Intelligence Is Revolutionising the Mining Industry." International Mining, August 23, 2023. im-mining.com/2023/08/23/how-artificial-intelligence-is-revolutionising-the-mining-industry.

[12] Smith, Ryan. "AI's Role in the Future of Forest Conservation." Network of Nature, March 19, 2024. networkofnature.org/blog/ai-s-role-in-the-future-of-forest-conservation.htm.

[13] Sharma, Gaurav. "How Multibillion Dollar Investments in AI Are Driving Oil and Gas Sector Innovation." Forbes, August 14, 2023. forbes.com/sites/gauravsharma/2023/08/14/how-multibillion-dollar-investments-in-ai-are-driving-oil-and-gas-sector-innovation.

[14] Patil, Amol. "The Good, the Bad, and the Awful of AI in Aerospace." Aerospace Manufacturing and Design, September 7, 2023. aerospacemanufacturinganddesign.com/article/the-good-the-bad-and-the-awful-of-ai-in-aerospace.

[15] dataforest.ai. "AI in Automotive: Pioneering Modern Vehicle Technology," May 27, 2024. dataforest.ai/blog/ai-in-automotive-transforming-the-automobile-industry.

[16] Santilli, Paul. "AI in Textile Manufacturing: Enhancing Efficiency and Sustainability- Strategic Consortium of Intelligence Professionals (SCIP)." Scip.org, June 6, 2024. scip.org/news/674390/AI-in-Textile-Manufacturing-Enhancing-Efficiency-and-Sustainability-.htm.

[17] Ao, Yu, Yunbo Li, Jiaye Gong, and Shaofan Li. "An Artificial Intelligence-Aided Design (AIAD) of Ship Hull Structures." Journal of Ocean Engineering and Science 8, no. 1 (January 2023): 15–32. doi.org/10.1016/j.joes.2021.11.003.

[18] Chemojo. "The Power of Prediction: How AI Is Transforming the Chemical Industry." Chemojo, March 29, 2024. chemojo.com/post/ai-in-chemical-industry.

[19] Dee, Catherine. "How AI Can Benefit the Retail Industry." Algolia, February 1, 2024. algolia.com/blog/ai/how-ai-can-benefit-the-retail-industry.

[20] Stefanini. "AI Transportation: Efficiency, Safety, and the Future," March 8, 2024. stefanini.com/en/insights/articles/artificial-intelligence-in-transportation-moving-faster.

[21] KPMG. "The Impact of Artificial Intelligence on the Insurance Industry," 2024. kpmg.com/us/en/articles/2024/impact-artificial-intelligence-insurance-industry.html.

[22] PublicisSapient. "5 Ways Artificial Intelligence Is Transforming the Restaurant Industry | Publicis Sapient." publicissapient.com, 2024. publicissapient.com/insights/ai-in-restaurants.

[23] Sun, Jane. "How Is AI Reshaping the Global Travel Experience?" World Economic Forum, December 19, 2023. weforum.org/agenda/2023/12/how-is-ai-reshaping-the-travel-tourism.

[24] Davenport, Thomas, and Randy Bean. "The Impact of Generative AI on Hollywood and Entertainment." MIT Sloan Management Review, June 19, 2023. sloanreview.mit.edu/article/the-impact-of-generative-ai-on-hollywood-and-entertainment.

[25] Bloomberg Law. "How Is AI Changing the Legal Profession?" May 23, 2024. pro.bloomberglaw.com/insights/technology/how-is-ai-changing-the-legal-profession/#is-ai-the-future-of-law.

[26] Davenport, Thomas, and Ravi Kalakota. "The Potential for Artificial Intelligence in Healthcare." Future Healthcare Journal 6, no. 2 (June 2019): 94–98. doi.org/10.7861/futurehosp.6-2-94.

[27] Deloitte. "How Artificial Intelligence Is Transforming the Financial Services Industry," 2023. deloitte.com/ng/en/services/risk-advisory/ services/how-artificial-intelligence-is-transforming-the-financial-services-industry. html.

[28] Chen, Claire. "AI Will Transform Teaching and Learning. Let's Get It Right." Stanford HAI. Stanford University, March 9, 2023. hai.stanford.edu/news/ai-will-transform-teaching-and-learning-lets-get-it-right.

[29] Mills, Steven, Greg Boison, Miguel Carrasco, and Nadim Abillama. "Unlocking the Value of AI-Powered Government." BCG Global, July 21, 2021. bcg.com/publications/2021/unlocking-value-ai-in-government.

[30] Delfino, Justin. "AI-Driven Research and Development: A Paradigm Shift in Innovation." Research & Development World, August 17, 2023. rdworldonline.com/ai-driven-research-and-development-a-paradigm-shift-in-innovation.

[31] The White House. "Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence." The White House, October 30, 2023. whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence.

[32] Duan, Jinhao, Hao Cheng, Shiqi Wang, Alex Zavalny, Chenan Wang, Renjing Xu, Bhavya Kailkhura, and Kaidi Xu. "Shifting Attention to Relevance: Towards the Predictive Uncertainty Quantification of Free-Form Large Language Models." Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (August 2024) 1: 5050-5063. aclanthology.org/2024.acl-long.276.

[33] Bartoldson, Brian R., James Diffenderfer, Konstantinos Parasyris, and Bhavya Kailkhura. "Adversarial Robustness Limits via Scaling-Law and Human-Alignment Studies." Proceedings of the 41st International Conference on Machine Learning 235 (July 8, 2024): 3046–72. proceedings.mlr.press/v235/bartoldson24a.html.

[34] Huang, Y., Sun, L., Wang, H., Wu, S., Zhang, Q., Li, Y., Gao, C., et al. (2024). "Position: TrustLLM: Trustworthiness in Large Language Models." Proceedings of the 41st International Conference on Machine Learning 235 (July 2024): 20166-20270. proceedings.mlr.press/v235/huang24x.html.

[35] Liu, Shusen, Donald W Loveland, Yong Han, and Bhavya Kailkhura. "Generative Counterfactual Introspection

for Explainable Deep Learning." Proceedings of the 2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP) (November 2019): 1-5. ieeexplore.ieee.org/document/8969491.

[36] Xu, Kaidi, Zhouxing Shi, Huan Zhang, Yihan Wang, Kai-Wei Chang, Minlie Huang, Bhavya Kailkhura, Xue Lin, and Cho-Jui Hsieh. "Automatic Perturbation Analysis for Scalable Certified Robustness and Beyond." Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS 2020) (October 2020): 1-24. doi.org/10.48550/arXiv.2002.12920.

[37] Richards, Luke E, Edward Raff, and Cynthia Matuszek. "Measuring Equality in Machine Learning Security Defenses: A Case Study in Speech Recognition." Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security (AISec '23) (November 2023): 161-171. dl.acm.org/doi/10.1145/3605764.3623911.

[38] Engel, Andrew, Zhichao Wang, Natalie Frank, Ioana Dumitriu, Sutanay Choudhury, Anand Sarwate, and Tony Chiang. "Faithful and Efficient Explanations for Neural Networks via Neural Tangent Kernel Surrogate Models." Proceedings of the 12th International Conference on Learning Representations (May 2024). openreview.net/forum?id=yKksu38BpM.

[39] Brown, Davis, Charles Godfrey, Cody Nizinski, Jonathan Tu, and Henry Kvinge. "Robustness of Edited Neural Networks." OpenReview, 2023. openreview.net/forum?id=JAjH6VANZ4.

[40] Kvinge, Henry, Tegan Emerson, Grayson Jorgenson, Scott Vasquez, Timothy Doster, and Jesse Lew. "In What Ways Are Deep Neural Networks Invariant and How Should We Measure This?" Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS 2022) 35: 32816-29. proceedings.neurips.cc/paper_files/paper/2022/hash/d36dfcdb14473a8526111c221660f2ab-Abstract-Conference.html.

[41] Toner, Helen, Jessica Ji, John Bansemer, and Lucy Lim. "Skating to Where the Puck Is Going | Anticipating and Managing Risks from Frontier AI Systems." Center for Security and Emerging Technology, November 16, 2023. cset.georgetown.edu/publication/skating-to-where-the-puck-is-going.

[42] Koch, James, Brenda Forland, Bruce Bernacki, Timothy Doster, and Tegan Emerson. "Data-Driven Invertible Neural Surrogates of Atmospheric Transmission." Proceedings of the 2024 IEEE International Geoscience and Remote Sensing Symposium (IGARSS) (July 2024): 6943-6947. ieeexplore.ieee.org/document/8969491.

[43] Field, Richard, Michael Smith, Joe Ingram, and Eva Domschot. "Modeling Correlated Features for Machine Learning Classification." Proceedings of the 17th U. S. National Congress on Computational Mechanics (July 2023):278. app.box.com/s/riuaiyw2tsdlt40jtzfamqiwdh55iqh7.

[44] Duersch, Jed A, and Thomas A Catanach. "Parsimonious Inference." arXiv.org, 2021. arxiv.org/abs/2103.02165.

[45] Soriano, Bruno S, Ki Sung Jung, Tarek Echekki, Jacqueline H Chen, and Mohammad Khalil. "Probabilistic Transfer Learning Methodology to Expedite High Fidelity Simulation of Reactive Flows." arXiv.org, 2024. arxiv.org/abs/2405.10944.

[46] Ries, Daniel, Joshua Michalenko, Tyler Ganter, Rashad Imad-Fayez Baiyasi, and Jason Adams. "Comparing the Quality of Neural Network Uncertainty Estimates for Classification Problems." Proceedings of the 21st IEEE International Conference on Machine Learning and Applications (ICMLA) (December 2022): 226-233. arxiv.org/abs/2308.05903

[47] Kegelmeyer, P, T.M. Shead, J Crussell, K Rodhouse, D Robinson, C Johnson, D Zage, et al. "Counter Adversarial Data Analytics." Proceedings of Graph Exploitation Symposium (July 2015). graphex.mit.edu/sites/default/files/documents/2015_Presentation_Kegelmeyer.pdf.

[48] Kegelmeyer, W. Philip, Wendt, Jeremy D, & Pinar, Ali. "An Example of Counter-Adversarial Community Detection Analysis." U.S. Department of Energy Office of Scientific and Technical Information, October 1, 2018. doi.org/10.2172/1481570.

[49] Field, Richard, and Michael Darling. "A Decision Theoretic Approach to Optimizing Machine Learning Decisions with Prediction Uncertainty." U.S. Department of Energy Office of Scientific and Technical Information, November 1, 2022. osti.gov/biblio/1899419.

[50] Matzen, L E, B C Howell, M Tuft, and K M Divis. “Transparent Risks: The Impact of the Specificity and Visual Encoding of Uncertainty on Decision Making.” Computer Graphics Forum 43, no. 3 (June 1, 2024). doi.org/10.1111/cgf.15094.

[51] Barrett, Anthony, Krystal Jackson, Evan Murphy, Nada Madkour, and Jessica Newman. “Benchmark Early and Red Team Often: A Framework for Assessing and Managing Dual-Use Hazards of AI Foundation Models,” May 2024. cltc.berkeley.edu/wp-content/uploads/2024/05/Dual-Use-Benchmark-Early-Red-Team-Often.pdf.

[52] Kang, Cecilia. “OpenAI’s Sam Altman Urges A.I. Regulation in Senate Hearing.” The New York Times, May 16, 2023, sec. Technology. nytimes.com/2023/05/16/technology/openai-altman-artificial-intelligence-regulation.html.

[53] The White House. “FACT SHEET: President Biden Issues Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence,” October 30, 2023. whitehouse.gov/briefing-room/statements-releases/2023/10/30/fact-sheet-president-biden-issues-executive-order-on-safe-secure-and-trustworthy-artificial-intelligence.

[54] Garza, Alejandro de la. “Inside the Nuclear Fusion Facility That Changed the World.” TIME, January 8, 2024. time.com/6344755/nuclear-fusion-nif.

[55] U.S. Department of Energy Office of Science, and National Nuclear Security Administration. Exascale Computing Project, n.d. exascaleproject.org.

[56] top500.org. “November 2024 | TOP500,” November 2024. top500.org/lists/top500/2024/11.

[57] Energy.gov. “Frontiers in Artificial Intelligence for Science, Security and Technology (FASST),” 2024. energy.gov/fasst.

[58] Future of Life Institute. “Pause Giant AI Experiments: An Open Letter.” Future of Life Institute, March 22, 2023. futureoflife.org/open-letter/pause-giant-ai-experiments.

[59] O’Brien, Matt, Kelvin Chan, and Thalia Beaty. “OpenAI Looks to Shift Away from Nonprofit Roots and Convert Itself to For-Profit Company.” AP News, September 26, 2024. apnews.com/article/chatgpt-openai-sam-altman-nonprofit-859bff5c19845f51796244e0072e2dfb.

[60] U.S. Department of Health and Human Services. “Surgeon General Issues New Advisory about Effects Social Media Use Has on Youth Mental Health,” May 23, 2023. hhs.gov/about/news/2023/05/23/surgeon-general-issues-new-advisory-about-effects-social-media-use-has-youth-mental-health.html.

Additional Reading

Carter, Jonathan, John Feddema, Doug Kothe, Rob Neely, Jason Pruet, and Rick Stevens. "Advanced Research Directions on AI for Science, Energy, and Security: Report on Summer 2022 Workshops." Argonne National Laboratory, May 2023.
publications.anl.gov/anlpubs/2023/06/182628.pdf.

Department of Energy. "Artificial Intelligence for Nuclear Deterrence Strategy 2023," 2023.
asc.llnl.gov/sites/asc/files/2024-03/na-114_2023_ai4nd_strategy_final_dist_20240117.pdf.

Bengio, Yoshua, Geoffrey Hinton, Andrew Yao, Dawn Song, Pieter Abbeel, Trevor Darrell, Yuval Noah Harari, et al. "Managing Extreme AI Risks amid Rapid Progress." Science 384, no. 6698 (May 20, 2024).
doi.org/10.1126/science.adn0117.

National Security Commission on Artificial Intelligence. "Chapter 11 is: Accelerating AI Innovation." Nscai.gov, 2024.
reports.nscai.gov/final-report/chapter-11.

Heim, Lennart, Tim Fist, Janet Egan, Sihao Huang, Stephen Zekany, Robert Trager, Michael A. Osborne, and Noa Zilberman. "Governing through the Cloud: The Intermediary Role of Compute Providers in AI Regulation," March 2024.
robots.ox.ac.uk/~mosb/public/pdf/6662/Heim%20et%20al.%20-%202024%20-%20Governing%20Through%20the%20Cloud%20The%20Intermediary%20Role.pdf.

UK Department for Science, Innovation, and Technology. "Capabilities and Risks from Frontier AI: A Discussion Paper on the Need for Further Research into AI Risk," October 2023.
assets.publishing.service.gov.uk/media/65395abae6c968000daa9b25/frontier-ai-capabilities-risks-report.pdf.

Dalrymple, David davidad, Joar Skalse, Yoshua Bengio, Stuart Russell, Max Tegmark, Sanjit Seshia, Steve Omohundro, et al. "Towards Guaranteed Safe AI: A Framework for Ensuring Robust and Reliable AI Systems." arXiv.org, July 8, 2024.
arxiv.org/abs/2405.06624.

Abbreviations and Acronyms

AGI	Artificial general intelligence
AI	Artificial intelligence
API	Application programming interface
ANL	Argonne National Laboratory
CBRN	Chemical, biological, radiological, and nuclear
CI	Cyberinfrastructure
DOE	Department of Energy
FASST	Frontiers in Artificial Intelligence for Science, Security and Technology
EO	Executive Order
HPC	High-performance computing
LANL	Los Alamos National Laboratory
LBNL	Lawrence Berkeley National Laboratory
LLM	Large language model
LLNL	Lawrence Livermore National Laboratory
ML	Machine learning
NIST	National Institute of Standards and Technology
PNNL	Pacific Northwest National Laboratory
R&D	Research and development
SNL	Sandia National Laboratories
UQ	Uncertainty quantification

LLNL Acknowledgments

Office of the Deputy Director for Science and Technology / **st.llnl.gov**

Data Science Institute / **data-science.llnl.gov**

Center for Advanced Signal and Image Sciences / **casis.llnl.gov**

University of California Livermore Collaboration Center – Event Hosts:
Camille Bibeau, Garren Weiss

Laboratory Director:
Kimberly Budil

Deputy Director for Science and Technology:
Patricia Falcone

Technical Advisors:
Brian Giera, Ruben Glatt, Michael Goldman, Cindy Gonzales, Felipe Leno da Silva, Brian Spears

Administrative Support:
Sira Neily, Kendall Luna, Jordan Coughenour

Production Staff:

Editors: Elisa Esme Abadi, Holly Auten

Designer: Mary J. Gines

Proofreader: Deanna Willis

Printers: Chris Brown, Kurt Johnson