Confidence-Building Measures for Artificial Intelligence: Workshop Proceedings

Sarah Shoker[1*], Andrew Reddie[2**], Sarah Barrington[2], Ruby Booth[3], Miles Brundage[1], Husanjot Chahal[1], Michael Depp[4], Bill Drexel[4], Ritwik Gupta[2], Marina Favaro[5], Jake Hecla[2], Alan Hickey[1], Margarita Konaev[6], Kirthi Kumar[2], Nathan Lambert[7], Andrew Lohn[6], Cullen O'Keefe[1], Nazneen Rajani[7], Michael Sellitto[5], Robert Trager[8], Leah Walker[2], Alexa Wehsener[9], Jessica Young[10]

[1]OpenAI, [2]University of California, Berkeley, [3]Berkeley Risk and Security Lab, [4]Center for a New American Security, [5]Anthropic, [6]Center for Security and Emerging Technology, [7]Hugging Face, [8]Centre for the Governance of AI, [9]Institute for Security and Technology, [10]Microsoft

August 2023

Abstract

Foundation models could eventually introduce several pathways for undermining state security: accidents, inadvertent escalation, unintentional conflict, the proliferation of weapons, and the interference with human diplomacy are just a few on a long list. The Confidence-Building Measures for Artificial Intelligence workshop hosted by the Geopolitics Team at OpenAI and the Berkeley Risk and Security Lab at the University of California brought together a multistakeholder group to think through the tools and strategies to mitigate the potential risks introduced by foundation models to international security. Originating in the Cold War, confidence-building measures (CBMs) are actions that reduce hostility, prevent conflict escalation, and improve trust between parties. The flexibility of CBMs make them a key instrument for navigating the rapid changes in the foundation model landscape. Participants identified the following CBMs that directly apply to foundation models and which are further explained in this conference proceedings: 1. crisis hotlines 2. incident sharing 3. model, transparency, and system cards 4. content provenance and watermarks 5. collaborative red teaming and table-top exercises and 6. dataset and evaluation sharing. Because most foundation model developers are non-government entities, many CBMs will need to involve a wider stakeholder community. These measures can be implemented either by AI labs or by relevant government actors.

All authors provided substantive contributions to the paper through sharing their ideas as participants in the workshop, writing the paper, and/or editorial feedback and direction. The first two authors are listed in order of contribution, and the remaining authors are listed alphabetically. Some workshop participants have chosen to remain anonymous. The claims in this paper do not represent the views of any author's organization. For questions about this paper, contact Sarah Shoker at sshoker@openai.com and Andrew Reddie at areddie@berkeley.edu.

*Significant contribution, including writing, providing detailed input for the paper, research, workshop organization, and setting the direction of the paper.

**Significant contribution, including providing detailed input for the paper, research, workshop organization, and setting the direction of the paper.

1 Introduction

Foundation models could eventually introduce several opportunities for undermining state security: accidents, inadvertent escalation, unintentional conflict,[1] the proliferation of weapons,[2] and the interference with human diplomacy are just a few on a long list.[3] Meanwhile, new defense and security actors continue to develop foundation model capabilities,[4] increasing the risk of an international crisis even further.

The *Confidence-Building Measures for Artificial Intelligence* workshop hosted by the Geopolitics Team at OpenAI and the Berkeley Risk and Security Lab (BRSL) at the University of California brought together participants from AI labs, government, academia, and civil society to propose tools and strategies to mitigate the potential risks introduced by foundation models to international security. By foundation models, we mean models that use vast amounts of data, self-supervision and deep learning methods which "can be adapted… to a wide range of downstream tasks."[5] The workshop included a mix of presentations and breakout groups, where participants had the opportunity to design possible confidence-building measures (CBMs). Together, participants identified the following CBMs that directly apply to foundation models:

- crisis hotlines
- incident sharing
- model, transparency, and system cards
- content provenance and watermarks
- collaborative red teaming exercises
- table-top exercises
- dataset and evaluation sharing

Popularized during the Cold War, CBMs represent "measures that address, prevent, or resolve uncertainties among states. Designed to prevent the escalation of hostilities and build mutual trust among erstwhile adversaries, CBMs can be formal or informal, unilateral, bilateral, or multilateral, [such as] military or political, and can be state-to-state or non-governmental."[6] Because states do not have perfect information about the capabilities or intentions of their allies and adversaries, formal and informal rules can establish predictability around state behavior, which in turn has the potential to reduce misunderstandings and miscommunications between state governments. This is in the interest of all parties.

1. In this context, accidents occur when AI systems malfunction. Inadvertent escalation happens due to inappropriate use of AI systems by leaders or operators that intensify situations. Unintentional conflict occurs when uncertainties in algorithm behavior hinder the ability of states to signal effectively to adversaries, potentially increasing the likelihood of conflict despite the ultimate intentions of involved states. Michael C. Horowitz and Lauren Kahn, "Leading in Artificial Intelligence through Confidence Building Measures," *The Washington Quarterly* 44, no. 4 (October 2021): 91–106, ISSN: 0163-660X, accessed July 17, 2023, https://doi.org/10.1080/0163660X.2021.2018794

2. *GPT-4 System Card*, March 2023.

3. Alexander Ward, Matt Berg, and Lawrence Ukenye, *Shaheen to Admin: Get Me the Black Sea Strategy*, https://www.politico.com/newsletters/national-security-daily/2023/03/21/shaheen-to-admin-get-me-the-black-sea-strategy-00088048, July 2023, accessed July 17, 2023.

4. *Palantir Artificial Intelligence Platform*, https://www.palantir.com/platforms/aip/, accessed July 17, 2023; *Donovan: AI-powered Decision-Making for Defense. | Scale AI*, https://scale.com/donovan, accessed July 17, 2023; Dan Milmo and Alex Hern, "UK to Invest £900m in Supercomputer in Bid to Build Own 'BritGPT'," *The Guardian*, March 2023, chap. Technology, ISSN: 0261-3077, accessed July 17, 2023; Jeffrey Ding and Jenny Xiao, "Recent Trends in China's Large Language Model Landscape," *Centre for the Governance of AI*, April 2023, accessed July 17, 2023.

5. Rishi Bommasani et al., *On the Opportunities and Risks of Foundation Models*, arXiv:2108.07258, July 2022, accessed July 17, 2023, https://doi.org/10.48550/arXiv.2108.07258, arXiv: 2108.07258 [cs].

6. *Confidence-Building Measures | Cross-Strait Security Initiative | CSIS*, https://www.csis.org/programs/international-security-program/isp-archives/asia-division/cross-strait-security-1, accessed July 17, 2023.

Historical examples of CBMs include direct call lines between countries to communicate during nuclear crises, reporting on weapon transfers between states, inviting observers to witness military exercises that an outside nation might otherwise construe as threatening, establishing clear "rules of the road" for how adversarial navies should interact on the high seas in peacetime, data exchanges on troop movements such as those mandated by the Treaty on Conventional Forces in Europe, or on-site monitoring of technology capabilities. In contrast to domestic or regional AI regulations that govern the relationship between companies and consumers, CBMs target and address the risks associated with state-to-state interactions by introducing predictability into a typically opaque international environment. While CBMs can target the prevention of a range of harms, workshop participants focused on CBMs that mitigate human rights abuses, the proliferation of unconventional weapons, and escalation due to misperceptions exacerbated by foundation models.

Defense strategies now routinely address the risks and opportunities associated with artificial intelligence, with some governments and think tanks calling explicitly for confidence-building measures.[7] Yet with the notable exception of the United Kingdom's Integrated Review Refresh 2023, most governments have not fully grappled with the implications of military AI, much less foundation models.[8] Though many existing defense documents do not directly target foundation models, governments can still fold the CBMs identified in these proceedings into existing AI commitments, such as the U.S Government's Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy.[9]

Building on the literature addressing the risks of AI to international security, this workshop focused on generating practical CBMs that apply to foundation models. The CBMs identified in these proceedings are not exhaustive or equally feasible in today's international climate. Where appropriate, we outline political and technical limitations that could interfere with the CBM's success.

1.1 A Note on Terminology: Foundation Models, Generative AI, and Large Language Models

For the sake of brevity, we use the term 'foundation model' to refer to both base and fine-tuned models, generative AI, and large language models. Where appropriate, we identify the specific type of AI model the CBM is meant to address. The terms foundation model, large language model, and generative AI are often used interchangeably, but there are significant, if imprecise, differences between these terms. As Helen Toner notes, these terms do not have "crisp boundaries…[but]...have emerged as attempts to point to a cluster of research directions and AI systems that have become especially noteworthy in recent years."[10]

Foundation models are built using deep learning and self-supervision learning methods and use vast amounts of data which, according to a 2022 paper by Rishi Bommasani et al. at Stanford University, "can be adapted (e.g. fine-tuned) to a wide range of downstream tasks."[11] The large amount of data and computational power used to train foundation models have led to impressive improvements across a variety of domains.[12]

While foundation models are often associated with generative AI applications like language and imagery (see below), these models can also be applied to domains such as robotics, human-machine interaction, reasoning, and sentiment analysis. On the other hand, generative AI is a narrower category of AI that includes models and algorithms capable of generating media. These models produce content like text, audio, imagery, and software code. Many public-facing models that are available today have already been fine-tuned on a foundation model. For example, ChatGPT models are fine-tuned on foundation models called GPT-3.5 and GPT-4, while Stability AI uses foundation models

7. *Chapter 4 - NSCAI Final Report*, technical report (National Security Commision on Artificial Intelligence), accessed July 17, 2023.
8. Page 56 Rishi Sunak, *Integrated Review Refresh 2023*, UK HM Government Report (HM Government, March 2023), 56
9. Bureau of Arms Control, *Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy - United States Department of State*, technical report (U.S. Department of State, February 2023), accessed July 17, 2023.
10. Helen Toner, *What Are Generative AI, Large Language Models, and Foundation Models?*, May 2023, accessed July 18, 2023.
11. Bommasani et al., *On the Opportunities and Risks of Foundation Models*.
12. *ibid.*

like StableLM to generate imagery.

1.2 Why do we need confidence-building measures for foundation models?

There is no shortage of historical crises where misperception or miscommunication led to military escalation that neither side wanted.[13] Misperception plays a prominent causal role in the bloodiest wars of the 20th century, whether that be in both World Wars, Cold War 'proxy' conflicts like Vietnam, Korea, and Afghanistan, or more recent 21st century conflicts like the Second Gulf War and Syrian Civil War. There are ample cases of militaries mistakenly targeting civilian planes and killing most or all civilians onboard,[14] and there are numerous historical false positives that only narrowly avoided nuclear exchange.[15]

The flexibility of CBMs make them a key instrument for navigating the rapid changes in the foundation model landscape. AI is a general purpose "enabling technology" rather than a military technology in and of itself.[16] For instance, current rule-making at the United Nations Convention on Certain Conventional Weapons (UN CCW) focuses on weapons identified by the forum,[17] which excludes many AI applications–such as generative AI–that are not obviously categorized as a 'weapon' but that can nevertheless influence the direction of international conflict. In particular, their non-binding, build-as-you-go nature allows the CBMs to grow in specificity as the technology necessarily evolves. This is essential, since it is not obvious what capabilities foundation models possess after they are trained and new capabilities are often revealed only after further red teaming and conducting safety evaluations. Though several benchmarks exist for assessing foundation models, they overwhelmingly point to rapid improvement in domain knowledge and deduction.[18] These capabilities are already associated with international security risks like providing information on the construction of conventional and unconventional weapons.[19]

CBMs do not overrule or subvert important efforts at fora like the United Nations and can act as an accompaniment to ongoing international regulatory discussions. CBMs are, however, uniquely equipped to target risks associated with foundation models due to the speed of their innovation and proliferation. In comparison to formal rules or international treaties, CBMs can lower coordination costs (such as time and money spent on bargaining) by reducing the number of negotiating parties involved in discussions. CBMs are often voluntary, which can incentivize participation from parties who are reluctant to risk the full weight of their national credibility on formal treaties. CBMs are more easily modified (and discarded).[20] CBMs can also 'start small' and build into formal rules, an especially useful feature in a low-trust international environment.[21]

Model performance and model safety are also separate research pursuits, meaning that the performance of foundation models can improve with little change to their safety profile. A large language model that can generate

13. Misperception continues to be a popular research area for scholars of military conflict, and some researchers suggest that the academic existence of international relations is fundamentally linked to managing problems related to information asymmetry and the anarchical conditions that make misperception possible.

14. Ron DePasquale, "Civilian Planes Shot Down: A Grim History," *The New York Times*, January 2020, chap. World, ISSN: 0362-4331, accessed July 18, 2023.

15. For a full list of nuclear false alarms, please visit compendium of events.*Close Calls with Nuclear Weapons*, technical report (Union of Concerned Scientists, January 2015), accessed July 18, 2023

16. Page 6 Iona Puscas, "Confidence-Building Measures for Artificial Intelligence: A Framing Paper," *United Nations Institute for Disarmament Research*, 2022, accessed July 17, 2023

17. *The Convention on Certain Conventional Weapons – UNODA*, technical report (United Nations Office for Diasmament Affairs), accessed July 18, 2023.

18. Dheeru Dua et al., *DROP: A Reading Comprehension Benchmark Requiring Discrete Reasoning Over Paragraphs*, arXiv:1903.00161, April 2019, accessed July 18, 2023, https://doi.org/10.48550/arXiv.1903.00161, arXiv: 1903.00161 [cs]; Dan Hendrycks, *Measuring Massive Multitask Language Understanding*, July 2023, accessed July 18, 2023; *Papers with Code - MMLU Benchmark (Multi-task Language Understanding)*, https://paperswithcode.com/sota/multi-task-language-understanding-on-mmlu, accessed July 18, 2023.

19. Page 12 *GPT-4 System Card*

20. Michael C. Horowitz, Lauren Kahn, and Casey Mahoney, "The Future of Military Applications of Artificial Intelligence: A Role for Confidence-Building Measures?," *Orbis* 64, no. 4 (January 2020): 528–543, ISSN: 0030-4387, accessed July 18, 2023, https://doi.org/10.1016/j.orbis.2020.08.003.

21. Horowitz and Kahn, "Leading in Artificial Intelligence through Confidence Building Measures."

information about nuclear physics is an example of a capability, while a large language model that refuses a user request to output specific details about bomb-building is an example of a safety mitigation. To date, AI labs have tackled the gap between model performance and safety by investing in a range of sociotechnical measures. Such measures include research into interpretability and alignment,[22] public disclosure of risks through system cards[23] and transparency notes,[24] delaying the release of models until sufficient safety mitigations have been implemented,[25] and open-sourcing evaluations[26] and provenance research.[27] Despite these efforts, the machine learning community is in general consensus that harm mitigations need further improvement to keep up with the rapidly increasing performance of LLMs.[28]

This landscape is further challenged by typical state behavior at the international level. States are often reluctant to engage in cooperative security agreements that require too much transparency into national capabilities. They are even less likely to place limits on the development of their own capabilities in the absence of any guarantees that their adversaries will do the same.[29] However, because performance and safety research are two different research streams, it is possible to coordinate on security while limiting availability of research into performance improvements. This unintended silver-lining is known to AI labs, which is why commercial labs are often willing to open-source safety research into evaluations and provenance technologies.

1.3 An Overview of CBMs for Foundation Models

Drawing from the list published by the United Nations Office of Disarmament Affairs, these proceedings organize CBMs under four categories: communication and coordination, observation and verification, cooperation and integration, and transparency.[30] These categories are not discrete; many CBMs can comfortably fit into more than one category.

22. Jeff Wu et al., *Recursively Summarizing Books with Human Feedback*, arXiv:2109.10862, September 2021, accessed July 18, 2023, https://doi.org/10.48550/arXiv.2109.10862, arXiv: 2109.10862 [cs]; Steven Bills et al., *Language Models Can Explain Neurons in Language Models*, OpenAI, May 2023; Yuntao Bai et al., *Constitutional AI: Harmlessness from AI Feedback*, arXiv:2212.08073, December 2022, accessed July 18, 2023, https://doi.org/10.48550/arXiv.2212.08073, arXiv: 2212.08073 [cs]; Nelson Elhage et al., *Toy Models of Superposition*, arXiv:2209.10652, September 2022, accessed July 18, 2023, https://doi.org/10.48550/arXiv.2209.10652, arXiv: 2209.10652 [cs]; Rohin Shah et al., *Goal Misgeneralization: Why Correct Specifications Aren't Enough For Correct Goals*, arXiv:2210.01790, November 2022, accessed July 18, 2023, https://doi.org/10.48550/arXiv.2210.01790, arXiv: 2210.01790 [cs]; Denny Zhou et al., *Least-to-Most Prompting Enables Complex Reasoning in Large Language Models*, arXiv:2205.10625, April 2023, accessed July 18, 2023, https://doi.org/10.48550/arXiv.2205.10625, arXiv: 2205.10625 [cs]; Daniel Ziegler et al., "Adversarial Training for High-Stakes Reliability," *Advances in Neural Information Processing Systems* 35 (December 2022): 9274–9286, accessed July 18, 2023; Michael K. Cohen, Marcus Hutter, and Michael A. Osborne, "Advanced Artificial Agents Intervene in the Provision of Reward," *AI Magazine* 43, no. 3 (2022): 282–293, ISSN: 2371-9621, accessed July 18, 2023, https://doi.org/10.1002/aaai.12064.

23. *GPT-4 System Card*.

24. ChrisHMSFT, *Transparency Note for Azure OpenAI - Azure Cognitive Services*, https://learn.microsoft.com/en-us/legal/cognitive-services/openai/transparency-note, May 2023, accessed July 18, 2023.

25. Page 19 *GPT-4 System Card* and *Microsoft Turing Academic Program (MS-TAP)*, accessed July 18, 2023

26. *Overview - C2PA*, https://c2pa.org/, accessed July 18, 2023; Paul England et al., *AMP: Authentication of Media via Provenance*, arXiv:2001.07886, June 2020, accessed July 18, 2023, https://doi.org/10.48550/arXiv.2001.07886, arXiv: 2001.07886 [cs, eess]; *Evals*, OpenAI, July 2023, accessed July 18, 2023.

27. For example, as part of its Content Authenticity Initiative (CAI), Adobe open-sourced JavaScript SDK and Rust SDK, which is "designed to let developers build functions displaying content credentials in browsers, or make custom desktop and mobile apps that can create, verify and display content credentials."Leigh Mc Gowran, *Adobe Launches Open-Source Tools to Tackle Visual Misinformation*, https://www.siliconrepublic.com/enterprise/adobe-digital-misinformation-cai-developer-tools, June 2022, accessed July 18, 2023

28. *Core Views on AI Safety: When, Why, What, and How*, https://www.anthropic.com/index/core-views-on-ai-safety, accessed July 28, 2023; Jan Leike, John Schulman, and Jeffrey Wu, *Our Approach to Alignment Research*, https://openai.com/blog/our-approach-to-alignment-research, August 2022, accessed July 28, 2023; Amelia Glaese et al., *Improving Alignment of Dialogue Agents via Targeted Human Judgements*, arXiv:2209.14375, September 2022, accessed July 28, 2023, https://doi.org/10.48550/arXiv.2209.14375, arXiv: 2209.14375 [cs].

29. For example, France's Defence Ethics Committee advised the continuation of research into autonomy and weapons systems, citing, among other reasons, the need to "counter enemy development of LAWS; and...to be able to defend ourselves against this type of weapon in the likely event of their use by an enemy State or terrorist group against our troops or population."*Opinion On The Integration Of Autonomy Into Lethal Weapon Systems*, technical report (Ministére Des Armées Defence Ethics Committee, April 2021), 5–6

30. *Repository of Military Confidence-Building Measures – UNODA*, accessed July 18, 2023.

Because most foundation model developers are non-government entities, many CBMs will need to involve a wider stakeholder community. These measures can be implemented either by AI labs or by relevant government actors.[31] Throughout the paper, we provide examples of adjacent technologies that have contributed to international crises, with the understanding that these examples can help us better anticipate the risks posed by foundation models, which are currently nascent or empirically unconfirmed.

1.4 Communication and Coordination

Communication and coordination CBMs reduce misunderstandings and misperceptions that, if left unaddressed, could escalate into conflict. The workshop identified two communication and coordination challenges that could be remedied using communication and coordination CBMs: misperceptions about authenticity of the content, and misperceptions concerning who authorized a decision.

First, on the topic of content authenticity, several workshop participants reiterated that foundation models and, specifically, generative AI, can be used to perpetuate 'truth decay,' or increased public distrust towards the information reported by political leaders and other experts. That distrust, in turn, complicates reporting on international events and crises.[32] For example, in March 2022, a widely circulated deepfake video on social media showed Ukrainian President Volodymyr Zelenskyy instructing soldiers to surrender to Russian forces.[33] Individuals may soon speak with *interactive* deepfakes, where the deepfake is both able to pause appropriately for the other speaker and use predictive modeling and synthetic audio to carry on a conversation.[34] And we could see the use of compositional deepfakes–not just one fake video or image, but many of them–released over time in between real events, to create a synthetic history that seems believable.[35]

Second, strong communication and coordination CBMs allow human actors to account for the ambiguity an AI injects into a system or team.[36] AI systems are often designed with the intention of supporting or augmenting human decision-making, making it challenging to disentangle the contributions of human operators. AI systems may also generate outputs or decisions that can be misinterpreted or misunderstood by human operators or other AI systems; in some cases the integration of AI in human-machine teams[37] can obfuscate whether AI was the (inadvertent, accidental, or intentional) cause of military escalation. A case in point is the 1988 tragedy of Iran Air Flight 655, which was targeted by an Aegis cruiser–the most sophisticated anti-aircraft weapon system at the time–on the order of the USS Vincennes, killing 290 civilians. The accident was blamed on a number of factors: the Aegis incorrectly identified the commercial airliner as a military aircraft; the commander of the Vincennes was characterized as being unnecessarily aggressive in a high-pressure atmosphere prone to misinterpretation; a nearby US navy ship, the USS Sides, had a human commander who correctly deduced that Iran Air Flight 655 was a civilian aircraft, but believed the Aegis's identification system to be technologically superior to his own human judgement and did not share his assessment with the Vincennes.[38] The Aegis radar system did eventually identify the Iran Air Flight 655 as a civilian

31. We opted to exclude a discussion on cyber risks from the scope of this paper since legal advances published in the Tallinn Manuals, NATO announcements on what counts as 'cyber war,' and norm setting at the UN Group of Governmental Experts on state behavior in cyberspace means that the topic deserves its own devoted forum.

32. Josh A. Goldstein et al., *Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations*, arXiv:2301.04246, January 2023, accessed July 18, 2023, https://doi.org/10.48550/arXiv.2301.04246, arXiv: 2301.04246 [cs]; Philip Oltermann, "European Politicians Duped into Deepfake Video Calls with Mayor of Kyiv," *The Guardian*, June 2022, chap. World news, ISSN: 0261-3077, accessed July 18, 2023.

33. Bobby Allyn, "Deepfake Video of Zelenskyy Could Be 'tip of the Iceberg' in Info War, Experts Warn," *NPR*, March 2022, chap. Technology, accessed July 18, 2023.

34. Eric Horvitz, "On the Horizon: Interactive and Compositional Deepfakes," in *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION* (November 2022), 653–661, accessed July 18, 2023, https://doi.org/10.1145/3536221.3558175, arXiv: 2209.01714 [cs].

35. *ibid.*

36. Horowitz and Kahn, "Leading in Artificial Intelligence through Confidence Building Measures."

37. We define 'human machine' team as "a relationship–one made up of at least three equally important elements: the human, the machine, and the interactions and interdependencies between them."*Building Trust in Human-Machine Teams*, https://www.brookings.edu/articles/building-trust-in-human-machine-teams/, accessed July 18, 2023

38. *H-020-1: USS Vincennes Tragedy*, http://public1.nhhcaws.local/content/history/nhhc/about-us/leadership/director/directors-corner/h-grams/h-gram-020/h-020-1-uss-vincennes-tragedy–.html, accessed July 28, 2023; *Formal Investigation into the Circumstances Surrounding the*

aircraft, but the human operators chose to accept the first reading. The Iran Air Flight 655 accident features many of the challenges that exist in today's human-machine teams: overtrust and a reluctance to contest the decisions made by the system, misunderstanding the threat due to existing geopolitical hostilities, and cherry-picking evidence to support one's interpretation of events. The introduction of AI to this atmosphere, which promises to increase the speed of targeting and analysis using a black-boxed technology, makes it even more necessary to identify communication pathways to prevent accidents.

Hotlines

The ability to interpret human intentions can become more challenging when communication integrates with or is supplanted by a non-human entity. Hotlines can assist with clarifying the 'who' or 'what' was responsible for military escalation, and for clarifying red lines[39] to avoid crossing them in the first place.[40] Workshop participants noted that competitor states could establish communication links to reduce friction during political crises, building on state-to-state hotlines that exist today for the management of military crises.

Despite their prominent role in mitigating nuclear crises, recent political events have underscored the reality that security norms will inform when parties use–or refuse–a phone call. This point was made especially evident during the February 2023 crisis involving a Chinese spy balloon traveling across the United States, and the subsequent refusal by the People's Liberation Army (PLA) to answer a hotline call from U.S Defense Secretary Lloyd Austin. Immediately following the crisis, researchers offered several explanations for the PLA's behavior that pointed to a discrepancy between how both military powers interpreted the threat landscape. Some stated that the PLA viewed CBMs and transparency as "disadvantageous" and a normalization "of increasingly brazen behavior." Another researcher stated that U.S military norms prize the avoidance of military escalation, while "[i]n the Chinese system, the impulse is to not be blamed for a mistake" or to be the person who reports the message to their political or military leaders.[41] It is worth noting that hotline usage becomes even more complicated in a world with three major military powers, where incentives could exist for one actor to exploit crisis communication between the two other states.

The successful use of hotlines may require that parties share common values about the risks of foundation models and a mutual belief that CBMs reduce the risk of unnecessary military escalations. States routinely disagree about the severity of threats and will pursue technologies to keep their own borders safe, even at the expense of global security. Other CBMs listed in this document, such as collaborative red teaming, emergency-response tabletop games, and incident sharing, can supply the necessary data for assessing the risk landscape while reinforcing the idea that CBMs will not undermine any single state's security. As an area of future study, participants recommended research on understanding policymaker perceptions about foundation model risks to international security and ensuring that incentives for participating in CBMs address country-specific social values.

Incident Sharing

Incident-sharing is a common practice across sectors where public safety is paramount, such as electric vehicles, cybersecurity, aviation, and healthcare. Information sharing about security incidents or 'near misses' is used to improve safety and reduce the likelihood of new accidents. With regards to autonomy in military systems, Michael Horowitz and Paul Scharre have previously suggested that an "'international autonomous incidents agreement' that focuses on military applications of autonomous systems, especially in the air and maritime environments... would reduce risks from accidental escalation by autonomous systems, as well as reduce ambiguity about the extent of human intention behind the behavior of autonomous systems".[42] This problem is documented in technology modernization

Downing of Iran Air Flight 655 on 3 July 1988, Investigation Report 93-FOI-0184 (U.S. Department of Defense, July 1988), 153.

39. Albert Wolf, *Backing Down: Why Red Lines Matter in Geopolitics*, https://mwi.westpoint.edu/geopolitical-costs-red-lines/, August 2016, accessed July 18, 2023.

40. For more information on how hotlines can clarify red lines, see: Bill Whitaker, *When Russian Hackers Targeted the U.S. Election Infrastructure*, https://www.cbsnews.com/news/when-russian-hackers-targeted-the-u-s-election-infrastructure/, July 2018, accessed July 18, 2023

41. Howard LaFranchi, "US-China Conundrum: Can Hotline Diplomacy Work If Trust Isn't a Goal?," *Christian Science Monitor*, March 2023, ISSN: 0882-7729, accessed July 18, 2023.

42. Michael C. Horowitz and Paul Scharre, *AI and International Stability: Risks and Confidence-Building Measures*, https://www.cnas.org/publications/reports/ai-and-international-stability-risks-and-confidence-building-measures, January 2021, accessed

for defense. For example, the introduction of touchscreen controls to the USS John S McCain, combined with crew confusion about the different settings associated with these controls, contributed to the largest maritime accident involving the US Navy in the last 40 years and left 10 sailors dead.[43]

Open-source AI incident-sharing initiatives already exist, such as the AI, Algorithmic, and Automation Incidents and Controversies (AIAAIC) and AI Incident Databases.[44] As of April 2023, these open-source databases primarily featured journalistic investigations, which sometimes include incidents on generative AI and international security, like the recent deepfake of the US Pentagon explosion.[45] Participants suggested a comparable database for international security incidents caused by foundation models, with a possible focus on unusual vulnerabilities and emergent model behaviors.

Workshop participants raised several questions that remain unresolved. Namely, it is unclear which model behaviors and misuses would qualify as an "incident," the incentives for parties to participate in an incident-sharing agreement, and how those parties can assure accurate reporting while respecting intellectual property rights and user privacy. A distinction might exist between new and dangerous model capabilities versus the large-scale misuse of the model. The former category could include behaviors linked to improvements of the model such as the ability to manipulate users or design a synthetic biochemical agent. The latter category could entail large-scale misuse campaigns, such as using models to create spam or disinformation.

Other industries resolve these challenges through data anonymization and encryption, trusted third-party agreements, access controls, NDAs, and security audits. Some types of incident sharing can leverage existing professional relationships between labs and could be as simple as hosting an informal meeting amongst parties. However, incident-sharing may require a multilateral entity that coordinates incident collection across multiple parties. Workshop participants noted that an AI incident-response entity could be analogous to existing Computer Emergency Response Teams (CERT) found in the international cybersecurity domain.[46]

2 Transparency

AI systems may produce unintended outcomes due to biases in training data, algorithmic errors, or unforeseen interactions with other systems. To name a few examples: foundation models used to summarize ISR data can introduce artifacts into the data that impacts a military response. Plausible outputs that are actually false, known as "hallucinations",[47] can be difficult to detect in a fast-paced and high-pressure military environment. Moreover, labeling practices can contribute to bias by privileging some worldviews over others, a serious risk for intelligence analysts conducting even routine tasks like report retrieval and summarization.[48] Compounding this problem is that models do not perform equally well across languages and it is seldom clear whose values should be reflected in model generations. Finally, prompt injection attacks, a type of data poisoning and security exploitation, can alter model

July 18, 2023.

43. *NTSB Accident Report on Fatal 2017 USS John McCain Collision off Singapore*, August 2019, chap. Documents, accessed July 18, 2023.

44. *AIAAIC*, https://www.aiaaic.org/home, accessed July 18, 2023; *AI Incidents Database*, https://partnershiponai.org/workstream/ai-incidents-database/, accessed July 18, 2023.

45. *Incident 543: Deepfake of Explosion Near US Military Administration Building Reportedly Causes Stock Dip*, https://incidentdatabase.ai/cite/543/, January 2020, accessed July 18, 2023.

46. For examples on national CERTs, please see: *US-CERT (United States Computer Emergency Readiness Team) - Glossary | CSRC*, https://csrc.nist.gov/glossary/term/us_cert, accessed July 18, 2023; *CERT-EU – Computer Emergency Response Team | European Union*, https://european-union.europa.eu/institutions-law-budget/institutions-and-bodies/search-all-eu-institutions-and-bodies/computer-emergency-response-team-eu-institutions-bodies-and-agencies-cert-eu_en, accessed July 18, 2023

47. Adrian Tam, *A Gentle Introduction to Hallucinations in Large Language Models*, June 2023, accessed July 18, 2023.

48. Scholars working within science and technology studies frequently note that the labeling of datasets reveal the political preferences and biases of the labellers, which has direct consequences for the model's performance. As Kate Crawford and Trevor Paglen note, "the automated interpretation of images is an inherently social and political project, rather than a purely technical one"Kate Crawford and Trevor Paglen, *Excavating AI*, https://excavating.ai, accessed July 19, 2023

outputs.[49] Prompt injection attacks are made easier when the adversary has access to the training data or model weights.

Some workshop participants also cited the problem of information overload. Even if accurate, too much information creates its own set of risks. States are often hesitant to escalate military activity because they are distrustful of their own predictions and intelligence about adversaries.[50] For example, machine learning could improve sensors to a point which renders the sea domain overly transparent and erodes the deterrent capacity of second strike forces.[51] In general, however, access to accurate information trends towards international stability.[52] To address the challenges posited above, workshop participants explored a variety of confidence-building measures. These are outlined, in brief, below.

2.1 Transparency Reports, Model and System Cards

System cards are documents that detail intended use cases, limitations, and the results of red teaming,[53] comparable to documentation practices found in industries such as aerospace, medicine, and pharmaceuticals.[54] In domestic settings, proponents of system cards argue that they can help policymakers better understand the capabilities and limitations of AI systems, informing oversight and regulation.[55] System cards do not require a third party to have access to the model itself, meaning that they can introduce transparency about capabilities while not revealing research details that would enable reverse-engineering of the model.

For foundation models used in defense domains, system cards should also include risks associated with human-machine interaction and overreliance, which can help outside observers interpret a system's behavior in the event of an accident or escalation. (It is not always possible to know who or what is responsible for a system's strange behavior.) For example, a 2021 UN Security Council report described a Turkish-made Kargu-2 drone in Libya as an instance where a lethal autonomous weapons system was deployed in violent conflict, a description that generated significant controversy in the international security community and highlighted the uncertainty involved with understanding the behavior of human-machine teams.[56]

For best effect, system cards should be readable and easily accessible. Many of today's system cards are found on code repository websites like Github, sites which tend not to be frequented by policymakers, and written in formats that those outside the field of machine learning can sometimes find inaccessible.

Like other measures found in these proceedings, there are limitations to model and system cards. Specifically, outside parties can experience difficulty verifying the results of model and system cards. Limitations can exist in non-adversarial and non-military contexts, too. Foundation models are often unavailable to third parties, with the

49. *Exploring Prompt Injection Attacks*, https://research.nccgroup.com/2022/12/05/exploring-prompt-injection-attacks/, December 2022, accessed July 18, 2023.

50. Glenn Herald Snyder, *Deterrence and Defense* (Princeton University Press, 1961), ISBN: 978-0-691-65209-2, accessed July 18, 2023.

51. James Johnson, "Artificial Intelligence, Drone Swarming and Escalation Risks in Future Warfare," *The RUSI Journal* 165, no. 2 (February 2020): 26–36, ISSN: 0307-1847, accessed July 18, 2023, https://doi.org/10.1080/03071847.2020.1752026.

52. Robert Jervis, "Cooperation Under the Security Dilemma," *World Politics* 30, no. 2 (January 1978): 167–214; James D. Fearon, "Rationalist Explanations for War," *International Organization* 49, no. 3 (1995): 379–414, JSTOR: 2706903; Charles A. Duelfer and Stephen Benedict Dyson, "Chronic Misperception and International Conflict: The US-Iraq Experience.," *International Security* 36, no. 4 (2011): 73–100, ISSN: 1531-4804.

53. Similarly, the Data Nutrition Project draws inspiration from nutrition labels found on food packaging. For an example from OpenAI, see *DallE-2 System Card*, https://github.com/openai/dalle-2-preview/blob/main/system-card.md, accessed July 18, 2023

54. Inioluwa Deborah Raji et al., "Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20 (New York, NY, USA: Association for Computing Machinery, January 2020), 33–44, ISBN: 978-1-4503-6936-7, accessed July 18, 2023, https://doi.org/10.1145/3351095.3372873; Margaret Mitchell et al., "Model Cards for Model Reporting," in *Proceedings of the Conference on Fairness, Accountability, and Transparency* (January 2019), 220–229, accessed July 19, 2023, https://doi.org/10.1145/3287560.3287596, arXiv: 1810.03993 [cs].

55. Mitchell et al., "Model Cards for Model Reporting."

56. Will Knight, "Autonomous Weapons Are Here, but the World Isn't Ready for Them," *Wired*: chap. tags, ISSN: 1059-1028, accessed July 18, 2023.

exception of base models made available by companies that open source their models. When foundation models are made publicly available, they are often refined through reinforcement learning from human feedback (RLHF), as seen in the InstructGPT models that power ChatGPT.[57] In general, publicly available fine-tuned models tend to be safer due to safety measures implemented after the base model has completed training.[58] Transparency reports, model and system cards often document the capabilities and limitations of the base model, making it difficult for third parties to replicate or otherwise validate the findings in these documents. This limitation is especially problematic in the context of international security, where adversaries may have several reasons to exaggerate or under-report the capabilities of their base models.

For this reason, model and system cards should be supported by other coordination activities, such as collaborative red teaming (explained in the section on 'Cooperation, Collaboration, and Integration.')

2.2 Observation and Verification

Parties can also agree to using empirical methods to observe and verify that actors are complying with agreements. These techniques do not normally guarantee full transparency, since states are reluctant to reveal the full scope of their military capabilities, as explained above. The Biological Weapons Convention (BWC), Chemical Weapons Convention (CWC), and the Treaty on the Non-Proliferation of Nuclear Weapons (NPT) all include mechanisms for third-party verification. Verification is also a key feature in agreements outside the UN. For example, the Open Skies Treaty allows signatories to fly observation aircraft to collect data on some military capabilities and activities.[59]

The success of verification is often dependent on the availability of detection and monitoring technologies. For this reason, detecting AI misuse in the international context has emphasized restrictions on hardware, since software can easily proliferate and evade monitoring. U.S. efforts have so far focused on constraining the supply of semiconductors and semiconductor manufacturing materials through export controls (though export controls are not necessarily confidence-building measures.)[60] These controls have mostly targeted two countries: Russia due to its invasion of Ukraine, and the PRC, which the 2022 U.S. National Security Strategy identifies as the "only competitor with the intent, and, increasingly, the capacity to reshape the international order."[61] However, algorithmic improvements, fine-tuning, the wide availability of consumer-grade LLM APIs, and open source alternatives mean that hardware controls are likely to be insufficient for preventing misuse. Moreover, while technology denial can constrain the range of choices available to any particular state, it does not address the problem of states' behavior with technologies already at their disposal.

2.3 Content Provenance and Watermarking

Content provenance and watermarking methods assist with the disclosure and detection of AI content and can reduce misperceptions by establishing norms around the use of generated content during international crises. Provenance and watermarking methods can improve traceability, alleviate concerns about the origin of the AI generated or edited content, and promote trust among parties. If properly vetted against adversarial manipulation, they can also help states use AI-generated products more confidently, knowing that the outcomes can be traced back to their source.

57. Ryan Lowe and Jan Leike, *Aligning Language Models to Follow Instructions*, https://openai.com/research/instruction-following, January 2022, accessed July 18, 2023.

58. Fine-tuning can also be used to make the model less safe, as actors could fine-tune base models on harmful or otherwise undesirable information. A recent example is the controversial release of 'GPT-4Chan,' where an open source model developed by Eleuther AI called GPT-J was finetuned on 4Chan forum posts. The model was hosted on the HuggingFace hub and even contained a model card before HuggingFace made the decision to restrict access and eventually remove the model. See: https://huggingface.co/ykilcher/gpt-4chan/discussions/1

59. "Treaty on Open Skies" (Helsinki, March 1992).

60. Laws that restrict the transfer of goods, technologies, and services to entities outside the country. Some export controls restrict this transfer to specific actors in another country ('end-user control') or restrict how the technology can be used ('end-use').

61. Page 8 Joe Biden, *2022 US National Security Strategy*, technical report (Washington: The White House, October 2022)

States can verify if the AI systems deployed by other parties adhere to agreed-upon guidelines or restrictions, making it easier to address any potential violations.

Content provenance is an ongoing and politically salient area of research, development, and adoption. For example, the Coalition for Content Provenance and Authenticity (C2PA), whose members include Adobe, Microsoft, Intel, BBC, Sony, and Truepic, is an industry-led initiative that develops technical standards for establishing the source and history of media content. Born to address the "prevalence of misleading information online," the coalition also offers a set of technical specifications for developers and guidelines to help users reason through the provenance of media content. As per the C2PA specifications, provenance methods can be split between "hard" and "soft" bindings, with the former including methods for applying unique identifiers to data assets and other cryptographic methods.[62] For instance, a C2PA manifest using cryptographically-bound provenance can include information about the origin of a piece of content (such as the AI model and version used to create it) and edits made to the content over time. A full survey of AI provenance methods is out of scope for this paper, but is well worth further research to determine which methods can be applied to improve state-to-state interactions and how these efforts complement each other.[63]

One of the most popular and readily-available AI disclosure methods in use today is "watermarking" (which the C2PA describes as a "soft" binding because they are more easily undermined in comparison to a "hard" binding.) Watermarking involves embedding low probability sequences of tokens into the outputs produced by AI systems, which can serve as a verification mechanism to confirm the authenticity and integrity of AI generations. Watermarks are traceable, meaning that they can enable parties to trace AI-generated outcomes back to their source system, thereby allowing stakeholders to identify which AI model was used and who was responsible for deploying it. However, watermarks are also accompanied by a severe restriction: they are not tamper-proof. For example, bad actors can use "paraphrasing attacks" to remove text watermarks, spoofing to infer hidden watermark signatures, or even add watermarks to authentic content.[64]

Because watermarking for large language models is a nascent area of research, the technique is mostly useful for photorealistic imagery, though watermarking for AI imagery also faces many limitations. Many AI-image generators that are publicly available today are already accompanied by watermarking methods, but these methods can be adversarially circumvented using post-processing methods on the image. For example, some watermarks for AI images can be removed through JPEG compression.[65] Due to such limitations, provenance tools should be frequently red teamed to verify their resilience against adversarial tampering, and C2PA provides security guidance to protect against attackers trying to tamper with provenance methods.

Watermarking is publicly popular, as demonstrated by a Vox Media survey that found that 78 percent of American adults believe that AI-generated media should be clearly disclosed.[66] Provenance and watermarks for imagery and audio feature prominently in the recent commitments made by AI companies to White House.[67] Despite its popularity with the U.S. public, the adoption of disclosure methods could also be contentious for both commercial and political reasons. First, small developers may not have the resources to invest in and apply provenance and watermarking technology. For this reason, open provenance standards and open sourcing AI detection technologies should be encouraged to help reduce the cost of security. Second, AI developers may also be reluctant to use watermarks on imagery for fear of alienating their consumers. Third, and as seen in the cybersecurity domain, states prefer certain technologies precisely because the malicious behavior is difficult to trace back to the belligerent party. States often

62. See section 2.4: *C2PA Security Considerations :: C2PA Specifications*, https://c2pa.org/specifications/specifications/1.0/security/Security_Considerat accessed July 18, 2023

63. For a list of disclosure methods, please see *PAI's Responsible Practices for Synthetic Media*, https://syntheticmedia.partnershiponai.org/, accessed July 28, 2023

64. Vinu Sankar Sadasivan et al., *Can AI-Generated Text Be Reliably Detected?*, arXiv:2303.11156, June 2023, accessed July 18, 2023, https://doi.org/10.48550/arXiv.2303.11156, arXiv: 2303.11156 [cs].

65. Zhengyuan Jiang, Jinghuai Zhang, and Neil Zhenqiang Gong, *Evading Watermark Based Detection of AI-Generated Content*, arXiv:2305.03807, May 2023, accessed July 18, 2023, https://doi.org/10.48550/arXiv.2305.03807, arXiv: 2305.03807 [cs].

66. Edwin H. Wong, *What Americans Are Really Excited about — and Scared of — When It Comes to AI*, https://www.voxmedia.com/2023/6/26/23769834/what-americans-are-really-excited-about-and-scared-of-when-it-comes-to-ai, June 2023, accessed July 28, 2023.

67. *Ensuring Safe, Secure, and Trustworthy AI*, technical report (Washington: The White House).

exploit technical and political ambiguities about "what counts" as an escalatory military behavior so that they can continue to engage in conflict below the threshold of war and avoid attribution.[68] Parties could exploit generative AI for the same reason, since it is currently unclear how the use of such models are interpreted by competitor states. While the proliferation of foundation models means that provenance and watermarking is unlikely to be applied evenly by all developers, states can commit–even unilaterally–to using such technologies in diplomatic and security activities.

2.4 Policies and Procedures

Rather than providing information about the models and systems that might be used, states could share information about the processes or procedures for assuring that they are safe. This could involve sharing baseline testing and best practices used to verify and validate AI-enabled systems. Since safety assurance is typically a mutually-aligned goal, some even envision that development of baseline testing techniques and procedures could become a collaborative effort among allies and adversaries.[69]

In addition to testing, publishing the policies and procedures for acquiring and approving AI-enabled systems can also provide confidence that systems are developed responsibly without divulging intellectual property. This could involve disclosing the minimum standards for performance such as the safety-integrity levels that exist for other safety-critical systems.[70] An option that reveals even less potentially sensitive information is to publicly name the responsible parties for approving the development, acquisition, and use of potentially worrisome capabilities. Even simply defining the capabilities that would require those extra approvals could help provide some clarity. DoD Directive 3000.09 does not satisfy all advocates, but it does make progress at providing clarity around some of these issues.

3 Cooperation, Collaboration, and Integration

Of course, many of the measures discussed above require AI labs and governments to collaborate and address the most proximate risks. Parties can coordinate security activities for the purpose of building trust and learning from one another. In higher trust environments, these activities can encourage transparency around military capabilities. In low trust environments, even simulation exercises can be difficult to organize.

3.1 Collaborative Red Teaming Exercises

Workshop participants advocated for collaborative red teaming, in (coincidental) alignment with the Biden Administration's recent announcement on responsible AI innovation, which featured a "public evaluation of generative AI systems."[71] Collaborative red-teaming in the United States is currently in development as a public transparency

68. Thomas Rid and Ben Buchanan, "Attributing Cyber Attacks," *The Journal of Strategic Studies* 38, nos. 1-2 (2015): 4–37, https://doi.org/10.1080/01402390.2014.977382.

69. Alexa Wehsener et al., *AI-NC3 Integration in an Adversarial Context: Strategic Stability Risks and Confidence Building Measures*, https://securityandtechnology.org/virtual-library/reports/ai-nc3-integration-in-an-adversarial-context-strategic-stability-risks-and-confidence-building-measures/, accessed July 18, 2023; Forrest E. Morgan et al., *Military Applications of Artificial Intelligence: Ethical Concerns in an Uncertain World*, technical report (RAND Corporation, April 2020), accessed July 18, 2023.

70. Andrew J. Lohn, *Estimating the Brittleness of AI: Safety Integrity Levels and the Need for Testing Out-Of-Distribution Performance*, arXiv:2009.00802, September 2020, accessed July 18, 2023, https://doi.org/10.48550/arXiv.2009.00802, arXiv: 2009.00802 [cs, stat].

71. In May 2023, The Biden Administration announced that major AI developers, including Anthropic, Google, Hugging https://www.overleaf.com/project/64b6a3eadcc06995e5dc3666Face, Microsoft, NVIDIA, OpenAI, and Stability AI, will participate in a public evaluation of generative AI systems at DEFCON 31's AI Village, with the aim of determining whether these mod-

and multistakeholder activity being held at DefCon 2023. The event includes participation by several companies, including Google, Hugging Face, Microsoft, NVIDIA, OpenAI, and Stability AI. Multilateral exercises similarly exist for cybersecurity, such as the annual Locked Shields exercise hosted by NATO's Cyber Defence Centre of Excellence (CCDCOE); in 2022, the red-team, blue-team exercise was attended by over two thousand cyber experts from thirty-two countries.[72] Unlike vulnerability discovery for cyber systems, red teaming foundation models often refers to capability discovery and require little background in machine learning. In turn, red team activities can improve emergency preparedness by exposing relevant stakeholders to the risks associated with foundation models.

3.2 Table-Top Exercises

Table-top exercises bring together stakeholders to simulate and discuss their responses to potential accidents or seemingly intractable solutions, improving crisis readiness and joint planning, and reducing the likelihood of misunderstandings during real-world conflicts. TTXs can also enhance coordination; states can develop a better understanding of each other's emergency procedures, identify areas where their response mechanisms or capabilities need improvement, and share best practices. TTXs between competitors, such as in track 2 diplomatic contexts, can improve the mutual understanding of intentions and surface risks or scenarios that might not have been considered.

Like red-teaming exercises, international fora can act as hosts for these events. The CCDCOE could integrate foundation models into Locked Shields to explore the prospect of cyber vulnerabilities, while the United Nations Institute for Disarmament Research (UNIDIR) could conduct a red-teaming exercise linked to the aims of the CCW discussions on autonomy in weapons systems. Because table-top exercises often serve as pedagogical tools, they can also be thought of as a 'training and education' CBM.

3.3 Dataset and Evaluation Sharing

Dataset sharing allows for the integration of safety standards across labs. Not to be confused with incident sharing, AI labs can collaborate on "refusals" by sharing datasets that focus on identifying and addressing safety or ethical concerns in AI-generated outputs. In the context of foundation models, "refusals" refer to instances where the AI system intentionally does not generate an output or refrains from providing a response to a user query due to safety or ethical concerns. This may occur when the requested output could potentially lead to harmful consequences, promote misinformation, or violate the ethical guidelines and policies set by the AI developers. Sharing such datasets can contribute to the development of more robust, aligned, and responsible AI systems.

In the area of international security, these datasets can contain information related to dual-use scientific information and that are legal to share across parties, such as in domains like chemical, biological, radiological, and nuclear science (CBRN). Red teaming has demonstrated that the interaction between LLMs and CBRN can introduce new proliferation pathways, potentially empowering non-state threat actors.[73] Sharing datasets allows AI labs to establish and integrate common benchmarks for evaluating the effectiveness of refusal mechanisms in LLMs. Datasets and evaluation sharing can also help to improve red teaming in small labs that do not have the resources to contract a large group of external experts.

There are limitations to dataset and evaluation sharing, especially as they relate to issues like national security. These limitations involve the regulation of dual-use items, the technical shortcomings of refusals, and expected future

els align with the Biden Administration's "Blueprint for an AI Bill of Rights" and "AI Risk Management Framework." See: The White House, *FACT SHEET: Biden-Harris Administration Announces New Actions to Promote Responsible AI Innovation That Protects Americans' Rights and Safety*, https://www.whitehouse.gov/briefing-room/statements-releases/2023/05/04/fact-sheet-biden-harris-administration-announces-new-actions-to-promote-responsible-ai-innovation-that-protects-americans-rights-and-safety/, May 2023, accessed July 18, 2023

72. *Locked Shields*, https://ccdcoe.org/exercises/locked-shields/, accessed July 18, 2023.

73. See page 12, 16: *GPT-4 System Card*

improvements in foundation models. Some scientific information is regulated to prevent the spread of information that can be used for weapons proliferation. For example, the U.S. International Traffic in Arms Regulation (ITAR) establishes controls on the import and export of technologies on the United States Munitions List (USML). The USML includes some dual-use technologies, and the regulations include rules surrounding the distribution of information about those technologies to foreign actors,[74] adding a layer of complexity to red teaming and the development of safety mitigations. As a result, many labs avoid this problem by red teaming information that is not controlled.

On a more general note, it is not yet clear if models will "discover" new biochemical compounds in the future and, if so, whether these discoveries may introduce new security vulnerabilities. Refusals target capabilities that have already been discovered, meaning they are a powerful but limited solution in the domain of biochemical discovery.

Despite these limitations, there are still benefits to sharing datasets that contain public, but sometimes difficult to find, information. Because information contained in these datasets are public, trialing dataset sharing today versus later is a comparatively lower risk endeavor, the main obstacle being the leveraging of scientific talent for capability discovery. In comparison, future data-sharing may be a higher risk activity if foundational models take on a greater role in biochemical discovery. And like all CBMs, the sharing of datasets demonstrates a commitment to transparency and responsible AI development, which can contribute to building trust among AI labs across countries, policymakers, and the broader public.

4 Conclusion

States are often reluctant to limit their technical capabilities. This is especially true in the face of heightened international competition and when confronting uncertainties around a new technology.[75] However, military interest in AI, and increasingly in foundation models and generative AI capabilities, has intensified the urgency of establishing an international code of conduct for state behavior, as the Political Declaration on Military Uses of Artificial Intelligence and Autonomy illustrates.[76] As Rebecca Hersman notes in her work on the potentially new escalation dynamics caused by emerging technologies, "[u]nlike traditional concepts of escalation, which suggest linear and somewhat predictable patterns from low-level crisis to all-out nuclear war, escalatory pathways in this new era of strategic competition will be less predictable."[77] While CBMs are not a complete solution for the international system's various insecurities, they do offer a set of options for reducing the likelihood of violent conflict caused by misinterpretation and miscalculation.

Though this workshop was primarily designed for thinking about solutions, participants identified a number of risks that could undermine the feasibility of CBM adoption. First, workshop attendees highlighted the information disparity between technologists and policymakers. The speed of politics–and diplomacy, in particular–often lags behind the speed of capability development, compounding the challenges of establishing appropriate CBMs in step with emerging tools. Policymakers may struggle to negotiate or provide assurances to their counterparts in other countries if they are unaware of the capabilities that exist within their borders.

Participants called for an increase in candid multistakeholder conversations to alleviate this problem. While the CBMs for AI workshop served as a space to discuss the impact of foundation models on international security, multistakeholder opportunities have predominantly been sporadic and reliant on the initiative of voluntary contributors

74. *International Traffic in Arms Regulations: U.S. Munitions List Categories I, II, and III*, https://www.federalregister.gov/documents/2020/01/23/2020-00574/international-traffic-in-arms-regulations-us-munitions-list-categories-i-ii-and-iii, January 2020, accessed July 18, 2023.

75. For example, France's Defence Ethics Committee advised the continuation of research into autonomy and weapons systems, citing, among other reasons, the need to "counter enemy development of LAWS; and…to be able to defend ourselves against this type of weapon in the likely event of their use by an enemy State or terrorist group against our troops or population." *Opinion On The Integration Of Autonomy Into Lethal Weapon Systems*

76. Horowitz and Kahn, "Leading in Artificial Intelligence through Confidence Building Measures."

77. Rebecca Hersman, *Wormhole Escalation in the New Nuclear Age*, https://tnsr.org/2020/07/wormhole-escalation-in-the-new-nuclear-age/, July 2020, accessed July 18, 2023.

to organize what are often time-consuming meetings.

Second, there are different options with respect to who would coordinate implementation and adoption of these CBMs, each with tradeoffs and drawbacks. For example, incident sharing demands not just funding, but also a reliable third party with sufficient staff to manage intake and ensure database quality. Participants suggested a variety of mechanisms to address this coordination issue, ranging from the addition of new offices within existing government agencies like the U.S. Office of Science and Technology Policy and parallel agencies in other countries, to forming international institutions that would oversee compliance and distribute benefits for participating states and labs. Two sandbox groups independently noted the oft-used and potentially problematic analogy of an 'IAEA for AI' as a comparable entity. Workshop participants suggested that states that abide by monitoring and verification norms could gain access to data pools, with initial access granted to testing and evaluation (T&E) infrastructure, followed by access to data and subsidies to support the T&E infrastructure of fledgling companies, further driving innovation and progress in the field and especially in safety research. However, for countries that are already data-rich, such as the United States and China, incentives focused on data sharing may be insufficient.

Third, it was unclear which incentives would encourage discussion and adoption of CBMs for different states. Disparities in military and technological capabilities among states may create resistance to CBMs, as some countries may believe that CBMs will disproportionately disadvantage them and data sharing incentives may be insufficient to overcome this perception. This belief may make a commitment to more intrusive monitoring and verification politically intractable in the near-term. There is an established literature in international relations that addresses the rational, normative, and psychological basis for participating in international agreements and arms control;[78] this literature provides a foundation for the development of incentives that could target the adoption of CBMs for foundation models.

Despite the challenge of transnational coordination, confidence-building measures that target foundation models are important for international stability. Some of the suggestions in this document are already being developed as consumer protections, meaning that much of the remaining work will be on persuading parties—both private and public—that the adoption of CBMs will be beneficial for international security, Equally promising are the growing calls for international coordination on AI by governments, technology companies, and civil society in the face of state-to-state competitive tensions.[79] These calls carve an opening for increased transnational dialogue between

78. Robert Jervis, "Arms Control, Stability, and Causes of War," *Political Science Quarterly* 108, no. 2 (1993): 239–253, ISSN: 0032-3195, accessed July 18, 2023, https://doi.org/10.2307/2152010, JSTOR: 2152010; Bernard Brodie, "On the Objectives of Arms Control," *International Security* 1, no. 1 (1976): 17–36, ISSN: 0162-2889, accessed July 18, 2023, https://doi.org/10.2307/2538574, JSTOR: 2538574; Hedley Bull, "Arms Control and World Order," *International Security* 1, no. 1 (1976): 3–16, ISSN: 0162-2889, accessed July 18, 2023, https://doi.org/10.2307/2538573, JSTOR: 2538573; Emanuel Adler, "The Emergence of Cooperation: National Epistemic Communities and the International Evolution of the Idea of Nuclear Arms Control," *International Organization* 46, no. 1 (1992): 101–145, ISSN: 0020-8183, accessed July 18, 2023, JSTOR: 2706953; Jon Brook Wolfsthal, "Why Arms Control?," *Daedalus* 149, no. 2 (2020): 101–115, ISSN: 0011-5266, accessed July 18, 2023, JSTOR: 48591315; Rose Gottemoeller, "Rethinking Nuclear Arms Control," *The Washington Quarterly* 43, no. 3 (July 2020): 139–159, ISSN: 0163-660X, accessed July 18, 2023, https://doi.org/10.1080/0163660X.2020.1813382; Richard K. Betts, "Systems for Peace or Causes of War? Collective Security, Arms Control, and the New Europe," *International Security* 17, no. 1 (1992): 5–43, ISSN: 0162-2889, accessed July 18, 2023, https://doi.org/10.2307/2539157, JSTOR: 2539157; Jeffrey Arthur Larsen, ed., *Arms Control: Cooperative Security in a Changing Environment* (Boulder, Colo: Lynne Rienner Publishers, 2002), ISBN: 978-1-58826-013-0; Abram Chayes, "An Inquiry into the Workings of Arms Control Agreements," *Harvard Law Review* 85, no. 5 (1972): 905–969, ISSN: 0017-811X, accessed July 18, 2023, https://doi.org/10.2307/1339933, JSTOR: 1339933; Kenneth L. Adelman, "Arms Control With and Without Agreements," *Foreign Affairs*, no. Winter 1984/85 (December 1984), ISSN: 0015-7120, accessed July 18, 2023; Steven E. Miller, "Politics over Promise: Domestic Impediments to Arms Control," *International Security* 8, no. 4 (1984): 67–90, ISSN: 0162-2889, accessed July 18, 2023, https://doi.org/10.2307/2538563, JSTOR: 2538563; Ivo H. Daalder, "The Future of Arms Control," *Survival* 34, no. 1 (March 1992): 51–73, ISSN: 0039-6338, accessed July 18, 2023, https://doi.org/10.1080/00396339208442630; Sarah E. Kreps, "The Institutional Design of Arms Control Agreements," *Foreign Policy Analysis* 14, no. 1 (January 2018): 127–147, ISSN: 1743-8586, accessed July 18, 2023, https://doi.org/10.1093/fpa/orw045; Andrew William Reddie, "Governing Insecurity: Institutional Design, Compliance, and Arms Control" (PhD diss., UC Berkeley, 2019), accessed July 18, 2023; Stephen Herzog, "After the Negotiations: Understanding Multilateral Nuclear Arms Control," *Yale Graduate School of Arts and Sciences Dissertations*, April 2021,

79. Kathleen Hicks, *Opinion | What the Pentagon Thinks About Artificial Intelligence*, https://www.politico.com/news/magazine/2023/06/15/pentagon-artificial-intelligence-china-00101751, June 2023, accessed July 18, 2023; Joseph Clark, *DOD Committed to Ethical Use of Artificial Intelligence*, https://www.defense.gov/News/News-Stories/Article/Article/3429864/dod-committed-to-ethical-use-of-artificial-intelligence/https%3A%2F%2Fwww.defense.gov%2FNews%2FNews-Stories%2FArticle%2FArticle%2F3429864%2Fdod-committed-to-ethical-use-of-artificial-intelligence%2F, June 2023, accessed July 18, 2023; Shen Weiduo, *OpenAI CEO Calls for Global Cooperation on AI*

civil societies and scientific communities. As non-government actors become increasingly responsible for steering technologies that have global ramifications, many of the sociotechnical solutions that reduce misperception will need to be implemented at the technical layer and in collaboration with private actors.

Acknowledgements

The authors would like to thank Wyatt Hoffman, Lauren Kahn, Philip Reiner, and Joel Parish for their valuable feedback on earlier versions of this manuscript.

Regulation, Says 'China Has Some of the Best AI Talent in the World' - Global Times, https://www.globaltimes.cn/page/202306/1292326.shtml, January 2023, accessed July 18, 2023.

References

Adelman, Kenneth L. "Arms Control With and Without Agreements." *Foreign Affairs*, no. Winter 1984/85 (December 1984). ISSN: 0015-7120, accessed July 18, 2023.

Adler, Emanuel. "The Emergence of Cooperation: National Epistemic Communities and the International Evolution of the Idea of Nuclear Arms Control." *International Organization* 46, no. 1 (1992): 101–145. ISSN: 0020-8183, accessed July 18, 2023. JSTOR: 2706953.

AIAAIC. https://www.aiaaic.org/home. Accessed July 18, 2023.

Allyn, Bobby. "Deepfake Video of Zelenskyy Could Be 'tip of the Iceberg' in Info War, Experts Warn." *NPR*, March 2022. Accessed July 18, 2023.

Core Views on AI Safety: When, Why, What, and How. https://www.anthropic.com/index/core-views-on-ai-safety. Accessed July 28, 2023.

Bai, Yuntao, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, et al. *Constitutional AI: Harmlessness from AI Feedback*, arXiv:2212.08073, December 2022. Accessed July 18, 2023. https://doi.org/10.48550/arXiv.2212.08073. arXiv: 2212.08073 [cs].

Betts, Richard K. "Systems for Peace or Causes of War? Collective Security, Arms Control, and the New Europe." *International Security* 17, no. 1 (1992): 5–43. ISSN: 0162-2889, accessed July 18, 2023. https://doi.org/10.2307/2539157. JSTOR: 2539157.

Biden, Joe. *2022 US National Security Strategy*. Technical report. Washington: The White House, October 2022.

Bills, Steven, Nick Cammarata, Dan Mossing, Henk Tillman, Leo Gao, Jeff Wu, and William Saunders. *Language Models Can Explain Neurons in Language Models*. OpenAI, May 2023.

Bommasani, Rishi, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, et al. *On the Opportunities and Risks of Foundation Models*, arXiv:2108.07258, July 2022. Accessed July 17, 2023. https://doi.org/10.48550/arXiv.2108.07258. arXiv: 2108.07258 [cs].

Brodie, Bernard. "On the Objectives of Arms Control." *International Security* 1, no. 1 (1976): 17–36. ISSN: 0162-2889, accessed July 18, 2023. https://doi.org/10.2307/2538574. JSTOR: 2538574.

Building Trust in Human-Machine Teams. https://www.brookings.edu/articles/building-trust-in-human-machine-teams/. Accessed July 18, 2023.

Bull, Hedley. "Arms Control and World Order." *International Security* 1, no. 1 (1976): 3–16. ISSN: 0162-2889, accessed July 18, 2023. https://doi.org/10.2307/2538573. JSTOR: 2538573.

Bureau of Arms Control. *Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy - United States Department of State*. Technical report. U.S. Department of State, February 2023. Accessed July 17, 2023.

C2PA Security Considerations :: C2PA Specifications. https://c2pa.org/specifications/specifications/1.0/security/Security_Cons
Accessed July 18, 2023.

Overview - C2PA. https://c2pa.org/. Accessed July 18, 2023.

Locked Shields. https://ccdcoe.org/exercises/locked-shields/. Accessed July 18, 2023.

CERT-EU – Computer Emergency Response Team | European Union. https://european-union.europa.eu/institutions-law-budget/institutions-and-bodies/search-all-eu-institutions-and-bodies/computer-emergency-response-team-eu-institutions-bodies-and-agencies-cert-eu_en. Accessed July 18, 2023.

Chapter 4 - NSCAI Final Report. Technical report. National Security Commision on Artificial Intelligence. Accessed July 17, 2023.

Chayes, Abram. "An Inquiry into the Workings of Arms Control Agreements." *Harvard Law Review* 85, no. 5 (1972): 905–969. ISSN: 0017-811X, accessed July 18, 2023. https://doi.org/10.2307/1339933. JSTOR: 1339933.

ChrisHMSFT. *Transparency Note for Azure OpenAI - Azure Cognitive Services*. https://learn.microsoft.com/en-us/legal/cognitive services/openai/transparency-note, May 2023. Accessed July 18, 2023.

Clark, Joseph. *DOD Committed to Ethical Use of Artificial Intelligence*. https://www.defense.gov/News/News-Stories/Article/Art committed-to-ethical-use-of-artificial-intelligence/https%3A%2F%2Fwww.defense.gov%2FNews%2FNews-Stories%2FAr committed-to-ethical-use-of-artificial-intelligence%2F, June 2023. Accessed July 18, 2023.

Close Calls with Nuclear Weapons. Technical report. Union of Concerned Scientists, January 2015. Accessed July 18, 2023.

Cohen, Michael K., Marcus Hutter, and Michael A. Osborne. "Advanced Artificial Agents Intervene in the Provision of Reward." *AI Magazine* 43, no. 3 (2022): 282–293. ISSN: 2371-9621, accessed July 18, 2023. https://doi.org/10.1002/aaai.

Confidence-Building Measures | Cross-Strait Security Initiative | CSIS. https://www.csis.org/programs/international-security-program/isp-archives/asia-division/cross-strait-security-1. Accessed July 17, 2023.

Crawford, Kate, and Trevor Paglen. *Excavating AI*. https://excavating.ai. Accessed July 19, 2023.

Daalder, Ivo H. "The Future of Arms Control." *Survival* 34, no. 1 (March 1992): 51–73. ISSN: 0039-6338, accessed July 18, 2023. https://doi.org/10.1080/00396339208442630.

DePasquale, Ron. "Civilian Planes Shot Down: A Grim History." *The New York Times*, January 2020. ISSN: 0362-4331, accessed July 18, 2023.

Ding, Jeffrey, and Jenny Xiao. "Recent Trends in China's Large Language Model Landscape." *Centre for the Governance of AI*, April 2023. Accessed July 17, 2023.

Dua, Dheeru, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. *DROP: A Reading Comprehension Benchmark Requiring Discrete Reasoning Over Paragraphs*, arXiv:1903.00161, April 2019. Accessed July 18, 2023. https://doi.org/10.48550/arXiv.1903.00161. arXiv: 1903.00161 [cs].

Duelfer, Charles A., and Stephen Benedict Dyson. "Chronic Misperception and International Conflict: The US-Iraq Experience." *International Security* 36, no. 4 (2011): 73–100. ISSN: 1531-4804.

Elhage, Nelson, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, et al. *Toy Models of Superposition*, arXiv:2209.10652, September 2022. Accessed July 18, 2023. https://doi.org/10.4 arXiv: 2209.10652 [cs].

England, Paul, Henrique S. Malvar, Eric Horvitz, Jack W. Stokes, Cédric Fournet, Rebecca Burke-Aguero, Amaury Chamayou, et al. *AMP: Authentication of Media via Provenance*, arXiv:2001.07886, June 2020. Accessed July 18, 2023. https://doi.org/10.48550/arXiv.2001.07886. arXiv: 2001.07886 [cs, eess].

Ensuring Safe, Secure, and Trustworthy AI. Technical report. Washington: The White House.

Evals. OpenAI, July 2023. Accessed July 18, 2023.

Fearon, James D. "Rationalist Explanations for War." *International Organization* 49, no. 3 (1995): 379–414. JSTOR: 2706903.

International Traffic in Arms Regulations: U.S. Munitions List Categories I, II, and III. https://www.federalregister.gov/documents 00574/international-traffic-in-arms-regulations-us-munitions-list-categories-i-ii-and-iii, January 2020. Accessed July 18, 2023.

Formal Investigation into the Circumstances Surrounding the Downing of Iran Air Flight 655 on 3 July 1988. Investigation Report 93-FOI-0184. U.S. Department of Defense, July 1988.

DallE-2 System Card. https://github.com/openai/dalle-2-preview/blob/main/system-card.md. Accessed July 18, 2023.

Glaese, Amelia, Nat McAleese, Maja Trębacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, et al. *Improving Alignment of Dialogue Agents via Targeted Human Judgements*, arXiv:2209.14375, September 2022. Accessed July 28, 2023. https://doi.org/10.48550/arXiv.2209.14375. arXiv: 2209.14375 [cs].

Goldstein, Josh A., Girish Sastry, Micah Musser, Renee DiResta, Matthew Gentzel, and Katerina Sedova. *Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations*, arXiv:2301.04246, January 2023. Accessed July 18, 2023. https://doi.org/10.48550/arXiv.2301.04246. arXiv: 2301.04246 [cs].

Gottemoeller, Rose. "Rethinking Nuclear Arms Control." *The Washington Quarterly* 43, no. 3 (July 2020): 139–159. ISSN: 0163-660X, accessed July 18, 2023. https://doi.org/10.1080/0163660X.2020.1813382.

Gowran, Leigh Mc. *Adobe Launches Open-Source Tools to Tackle Visual Misinformation*. https://www.siliconrepublic.com/enterpr digital-misinformation-cai-developer-tools, June 2022. Accessed July 18, 2023.

GPT-4 System Card, March 2023.

H-020-1: USS Vincennes Tragedy. http://public1.nhhcaws.local/content/history/nhhc/about-us/leadership/director/directors corner/h-grams/h-gram-020/h-020-1-uss-vincennes-tragedy–.html. Accessed July 28, 2023.

Hendrycks, Dan. *Measuring Massive Multitask Language Understanding*, July 2023. Accessed July 18, 2023.

Hersman, Rebecca. *Wormhole Escalation in the New Nuclear Age*. https://tnsr.org/2020/07/wormhole-escalation-in-the-new-nuclear-age/, July 2020. Accessed July 18, 2023.

Herzog, Stephen. "After the Negotiations: Understanding Multilateral Nuclear Arms Control." *Yale Graduate School of Arts and Sciences Dissertations*, April 2021.

Hicks, Kathleen. *Opinion | What the Pentagon Thinks About Artificial Intelligence*. https://www.politico.com/news/magazine/20 artificial-intelligence-china-00101751, June 2023. Accessed July 18, 2023.

Horowitz, Michael C., and Lauren Kahn. "Leading in Artificial Intelligence through Confidence Building Measures." *The Washington Quarterly* 44, no. 4 (October 2021): 91–106. ISSN: 0163-660X, accessed July 17, 2023. https://doi.org/10.

Horowitz, Michael C., Lauren Kahn, and Casey Mahoney. "The Future of Military Applications of Artificial Intelligence: A Role for Confidence-Building Measures?" *Orbis* 64, no. 4 (January 2020): 528–543. ISSN: 0030-4387, accessed July 18, 2023. https://doi.org/10.1016/j.orbis.2020.08.003.

Horowitz, Michael C., and Paul Scharre. *AI and International Stability: Risks and Confidence-Building Measures*. https://www.cnas.org/publications/reports/ai-and-international-stability-risks-and-confidence-building-measures, January 2021. Accessed July 18, 2023.

Horvitz, Eric. "On the Horizon: Interactive and Compositional Deepfakes." In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION*, 653–661. November 2022. Accessed July 18, 2023. https://doi.org/10.1145/3536221.3558 arXiv: 2209.01714 [cs].

House, The White. *FACT SHEET: Biden-Harris Administration Announces New Actions to Promote Responsible AI Innovation That Protects Americans' Rights and Safety*. https://www.whitehouse.gov/briefing-room/statements-releases/2023/05/04/fact-sheet-biden-harris-administration-announces-new-actions-to-promote-responsible-ai-innovation-that-protects-americans-rights-and-safety/, May 2023. Accessed July 18, 2023.

Incident 543: Deepfake of Explosion Near US Military Administration Building Reportedly Causes Stock Dip. https://incidentdataba January 2020. Accessed July 18, 2023.

Jervis, Robert. "Arms Control, Stability, and Causes of War." *Political Science Quarterly* 108, no. 2 (1993): 239–253. ISSN: 0032-3195, accessed July 18, 2023. https://doi.org/10.2307/2152010. JSTOR: 2152010.

———. "Cooperation Under the Security Dilemma." *World Politics* 30, no. 2 (January 1978): 167–214.

Jiang, Zhengyuan, Jinghuai Zhang, and Neil Zhenqiang Gong. *Evading Watermark Based Detection of AI-Generated Content*, arXiv:2305.03807, May 2023. Accessed July 18, 2023. https://doi.org/10.48550/arXiv.2305.03807. arXiv: 2305.03807 [cs].

Johnson, James. "Artificial Intelligence, Drone Swarming and Escalation Risks in Future Warfare." *The RUSI Journal* 165, no. 2 (February 2020): 26–36. ISSN: 0307-1847, accessed July 18, 2023. https://doi.org/10.1080/03071847.2020.1

Knight, Will. "Autonomous Weapons Are Here, but the World Isn't Ready for Them." *Wired*. ISSN: 1059-1028, accessed July 18, 2023.

Kreps, Sarah E. "The Institutional Design of Arms Control Agreements." *Foreign Policy Analysis* 14, no. 1 (January 2018): 127–147. ISSN: 1743-8586, accessed July 18, 2023. https://doi.org/10.1093/fpa/orw045.

LaFranchi, Howard. "US-China Conundrum: Can Hotline Diplomacy Work If Trust Isn't a Goal?" *Christian Science Monitor*, March 2023. ISSN: 0882-7729, accessed July 18, 2023.

Larsen, Jeffrey Arthur, ed. *Arms Control: Cooperative Security in a Changing Environment*. Boulder, Colo: Lynne Rienner Publishers, 2002. ISBN: 978-1-58826-013-0.

Leike, Jan, John Schulman, and Jeffrey Wu. *Our Approach to Alignment Research*. https://openai.com/blog/our-approach-to-alignment-research, August 2022. Accessed July 28, 2023.

Lohn, Andrew J. *Estimating the Brittleness of AI: Safety Integrity Levels and the Need for Testing Out-Of-Distribution Performance*, arXiv:2009.00802, September 2020. Accessed July 18, 2023. https://doi.org/10.48550/arXiv.2009.00802. arXiv: 2009.00802 [cs, stat].

Lowe, Ryan, and Jan Leike. *Aligning Language Models to Follow Instructions*. https://openai.com/research/instruction-following, January 2022. Accessed July 18, 2023.

Microsoft Turing Academic Program (MS-TAP). Accessed July 18, 2023.

Miller, Steven E. "Politics over Promise: Domestic Impediments to Arms Control." *International Security* 8, no. 4 (1984): 67–90. ISSN: 0162-2889, accessed July 18, 2023. https://doi.org/10.2307/2538563. JSTOR: 2538563.

Milmo, Dan, and Alex Hern. "UK to Invest £900m in Supercomputer in Bid to Build Own 'BritGPT'." *The Guardian*, March 2023. ISSN: 0261-3077, accessed July 17, 2023.

Mitchell, Margaret, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. "Model Cards for Model Reporting." In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 220–229. January 2019. Accessed July 19, 2023. https://doi.org/10.1145/3287 arXiv: 1810.03993 [cs].

Morgan, Forrest E., Benjamin Boudreaux, Andrew J. Lohn, Mark Ashby, Christian Curriden, Kelly Klima, and Derek Grossman. *Military Applications of Artificial Intelligence: Ethical Concerns in an Uncertain World*. Technical report. RAND Corporation, April 2020. Accessed July 18, 2023.

Exploring Prompt Injection Attacks. https://research.nccgroup.com/2022/12/05/exploring-prompt-injection-attacks/, December 2022. Accessed July 18, 2023.

Oltermann, Philip. "European Politicians Duped into Deepfake Video Calls with Mayor of Kyiv." *The Guardian*, June 2022. ISSN: 0261-3077, accessed July 18, 2023.

Opinion On The Integration Of Autonomy Into Lethal Weapon Systems. Technical report. Ministére Des Armées Defence Ethics Committee, April 2021.

Palantir Artificial Intelligence Platform. https://www.palantir.com/platforms/aip/. Accessed July 17, 2023.

Papers with Code - MMLU Benchmark (Multi-task Language Understanding). https://paperswithcode.com/sota/multi-task-language-understanding-on-mmlu. Accessed July 18, 2023.

AI Incidents Database. https://partnershiponai.org/workstream/ai-incidents-database/. Accessed July 18, 2023.

PAI's Responsible Practices for Synthetic Media. https://syntheticmedia.partnershiponai.org/. Accessed July 28, 2023.

Puscas, Iona. "Confidence-Building Measures for Artificial Intelligence: A Framing Paper." *United Nations Institute for Disarmament Research*, 2022. Accessed July 17, 2023.

Raji, Inioluwa Deborah, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. "Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing." In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 33–44. FAT* '20. New York, NY, USA: Association for Computing Machinery, January 2020. ISBN: 978-1-4503-6936-7, accessed July 18, 2023. https://doi.org/10.1145/3351095.3372873.

Reddie, Andrew William. "Governing Insecurity: Institutional Design, Compliance, and Arms Control." PhD diss., UC Berkeley, 2019. Accessed July 18, 2023.

Repository of Military Confidence-Building Measures – UNODA. Accessed July 18, 2023.

Rid, Thomas, and Ben Buchanan. "Attributing Cyber Attacks." *The Journal of Strategic Studies* 38, nos. 1-2 (2015): 4–37. https://doi.org/10.1080/01402390.2014.977382.

Sadasivan, Vinu Sankar, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. *Can AI-Generated Text Be Reliably Detected?*, arXiv:2303.11156, June 2023. Accessed July 18, 2023. https://doi.org/10.48550/arXiv.2303.11 arXiv: 2303.11156 [cs].

Donovan: AI-powered Decision-Making for Defense. | Scale AI. https://scale.com/donovan. Accessed July 17, 2023.

Shah, Rohin, Vikrant Varma, Ramana Kumar, Mary Phuong, Victoria Krakovna, Jonathan Uesato, and Zac Kenton. *Goal Misgeneralization: Why Correct Specifications Aren't Enough For Correct Goals*, arXiv:2210.01790, November 2022. Accessed July 18, 2023. https://doi.org/10.48550/arXiv.2210.01790. arXiv: 2210.01790 [cs].

Snyder, Glenn Herald. *Deterrence and Defense*. Princeton University Press, 1961. ISBN: 978-0-691-65209-2, accessed July 18, 2023.

Sunak, Rishi. *Integrated Review Refresh 2023*. UK HM Government Report. HM Government, March 2023.

Tam, Adrian. *A Gentle Introduction to Hallucinations in Large Language Models*, June 2023. Accessed July 18, 2023.

The Convention on Certain Conventional Weapons – UNODA. Technical report. United Nations Office for Diasmament Affairs. Accessed July 18, 2023.

Toner, Helen. *What Are Generative AI, Large Language Models, and Foundation Models?*, May 2023. Accessed July 18, 2023.

"Treaty on Open Skies." Helsinki, March 1992.

US-CERT (United States Computer Emergency Readiness Team) - Glossary | CSRC. https://csrc.nist.gov/glossary/term/us_cert. Accessed July 18, 2023.

NTSB Accident Report on Fatal 2017 USS John McCain Collision off Singapore, August 2019. Accessed July 18, 2023.

Ward, Alexander, Matt Berg, and Lawrence Ukenye. *Shaheen to Admin: Get Me the Black Sea Strategy*. https://www.politico.com/ security-daily/2023/03/21/shaheen-to-admin-get-me-the-black-sea-strategy-00088048, July 2023. Accessed July 17, 2023.

Wehsener, Alexa, Andrew W. Reddie, Leah Walker, and Philip Reiner. *AI-NC3 Integration in an Adversarial Context: Strategic Stability Risks and Confidence Building Measures*. https://securityandtechnology.org/virtual-library/reports/ai-nc3-integration-in-an-adversarial-context-strategic-stability-risks-and-confidence-building-measures/. Accessed July 18, 2023.

Weiduo, Shen. *OpenAI CEO Calls for Global Cooperation on AI Regulation, Says 'China Has Some of the Best AI Talent in the World' - Global Times*. https://www.globaltimes.cn/page/202306/1292326.shtml, January 2023. Accessed July 18, 2023.

Whitaker, Bill. *When Russian Hackers Targeted the U.S. Election Infrastructure*. https://www.cbsnews.com/news/when-russian-hackers-targeted-the-u-s-election-infrastructure/, July 2018. Accessed July 18, 2023.

Wolf, Albert. *Backing Down: Why Red Lines Matter in Geopolitics*. https://mwi.westpoint.edu/geopolitical-costs-red-lines/, August 2016. Accessed July 18, 2023.

Wolfsthal, Jon Brook. "Why Arms Control?" *Daedalus* 149, no. 2 (2020): 101–115. ISSN: 0011-5266, accessed July 18, 2023. JSTOR: 48591315.

Wong, Edwin H. *What Americans Are Really Excited about — and Scared of — When It Comes to AI*. https://www.voxmedia.com/2 americans-are-really-excited-about-and-scared-of-when-it-comes-to-ai, June 2023. Accessed July 28, 2023.

Wu, Jeff, Long Ouyang, Daniel M. Ziegler, Nisan Stiennon, Ryan Lowe, Jan Leike, and Paul Christiano. *Recursively Summarizing Books with Human Feedback*, arXiv:2109.10862, September 2021. Accessed July 18, 2023. https://doi.org/10.48550/arXiv.2109.10862. arXiv: 2109.10862 [cs].

Zhou, Denny, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, et al. *Least-to-Most Prompting Enables Complex Reasoning in Large Language Models*, arXiv:2205.10625, April 2023. Accessed July 18, 2023. https://doi.org/10.48550/arXiv.2205.10625. arXiv: 2205.10625 [cs].

Ziegler, Daniel, Seraphina Nix, Lawrence Chan, Tim Bauman, Peter Schmidt-Nielsen, Tao Lin, Adam Scherlis, et al. "Adversarial Training for High-Stakes Reliability." *Advances in Neural Information Processing Systems* 35 (December 2022): 9274–9286. Accessed July 18, 2023.