Introduction:

AI is a disruptive and powerful technology that is being rapidly integrated with U.S. military infrastructure. Through an examination of some of the likely AI integrations with nuclear weapon systems, I will explore several of the destabilizing threats that are likely to be introduced to the US nuclear arsenal.

I will explore through the lens of two approaches: First, I will focus on well-established flaws inherent to AI and second, I will explore the hazards stemming from the overwhelming competitive pressures that may lead to the hasty deployment of systems that are under-tested and consequently unsafe.

In researching this article, I found little specific information concerning the technical impact of AI on nuclear weapons. There were a number of well written articles that did analyze AI's effect on the nuclear balance, but relatively few focus on the technological impact, and almost none that provided a high level of detail and specificity (with the notable exception of Peter Rautenbach's indispensable research linked here). Peter's report has been an invaluable source for this article.

Defining AI:

Artificial intelligence as a term has been broadly applied to the very diverse ways in which computers accomplish tasks that historically have required human thinking. The recent massive advances in the capabilities of AI to perform more and more intelligent tasks has been powered by machine learning where algorithms enable computers to do tasks without explicit instruction. AI's essence is the ability to spot patterns and make predictions. The latest breakthroughs in AI have been concentrated in a specific form of machine learning known as deep learning. Deep learning uses neural nets to better understand very large amounts of data which in turn can increase performance on complex tasks.

The recent concerns about AI launched by chatGPT's release center on a specific form of AI that uses deep learning to create models that predict the next word in a sentence. These models are known as large language models (LLMs). The latest LLMs have been trained on vast swathes of the internet, enabling them to write with intelligence that eerily mimics humans in many domains. LLMs are now being incorporated into the military: "Both the U.S. Marine Corps and the U.S. Air Force are experimenting with LLMs, using them for war games, military planning, and basic administrative tasks. Palantir, a company that develops information technology for the DOD, has created a product that uses LLMs to manage military operations". Unfortunately researchers have found disturbing characteristics in wargaming simulations:

"...LLMs chose escalation and exhibited a preference toward arms races, conflict, and even the use of nuclear weapons".[1]

While LLMs are based on language, their capabilities have been integrated with computer vision, the ability of a computer to understand an image. OpenAI's latest LLM, GPT-4, was combined with vision to produce GPT-4 with vision (GPT-4V) which "...enables users to instruct GPT-4 to analyze image inputs provided by the user…".[2]

Machine learning (whether LLM powered or other kinds) can be integrated with a variety of nuclear architecture systems and is particularly likely to be applied to two types of systems: early warning and decision support. It is in these domains where AI's ability to "recognize complex patterns in pictures, text, sounds, and other data to produce accurate insights and predictions"[3] will undoubtedly be leveraged.

Nuclear Infrastructure and the push to incorporate AI

The term applied to the systems that support and coordinate nuclear weapons is nuclear command, control, and communication or NC3. Historical examples of errors in NC3 architecture nearly causing accidental nuclear conflict abound as depicted on the linked timeline. The historical false alarms that have been sounded may inspire calls for modernizing our NC3 by updating it with machine learning across the estimated 39% of NC3 systems that could be improved from it.[4] While details of these integrations are not public, there are indications that the modernization is being contemplated: the former commander of U.S. Strategic Command, General Hyten, acknowledged that "...AI can play an important part" in NC3 systems.[5] The sudden level of military interest in incorporating AI is enormous: "the potential value of AI-related federal contracts increased by almost 1,200%, from $355 million in the period leading up to August 2022, to $4.6 billion in the period leading up to August 2023. This increase was almost entirely driven by the Department of Defense (DoD)"[6].

The Uncertainties Generated by Competitive Pressure

The present international competitive pressure for AI superiority echoes the international rivalry, first with Germany and later with the Soviet Union that led to the development of nuclear weapons. Beyond the basic desire to enhance and modernize the capabilities of U.S. systems, is the pressure exerted by a competitive desire to stay ahead of rival nuclear nations, most

---

[1] https://www.foreignaffairs.com/united-states/why-military-cant-trust-ai#:~:text=Despite%20these%20differences%2C%20we%20found,the%20use%20of%20nuclear%20weapons.
[2] https://openai.com/index/gpt-4v-system-card/
[3] https://aws.amazon.com/what-is/deep-learning/#:~:text=Deep%20learning%20is%20a%20method,produce%20accurate%20insights%20and%20predictions.
[4] https://www.c4isrnet.com/thought-leadership/2020/04/30/when-machine-learning-comes-to-nuclear-communication-systems/
[5] https://breakingdefense.com/2018/04/stratcoms-hyten-on-b-21-colombia-class-nc3/
[6] https://time.com/6961317/ai-artificial-intelligence-us-military-spending/

notably China and Russia. These are key drivers that have created a widespread belief that developing and deploying AI for military purposes is urgent. As Adm. Christopher Grady, vice chairman of the Joint Chiefs of Staff stated regarding AI: "Whether you want to call it a race or not, it certainly is… Both of us have recognized that this will be a very critical element of the future battlefield. China's working on it as hard as we are"[7]. The obvious danger of an arms race driven by a new and uncertain technology is that fear, secrecy, and hyperbole may lead to short cuts and ill advised decisions.

Historically, the false perception that the U.S. was facing a "missile gap" with the Soviet Union (when in fact they had a total of 4 ICBMs[8]) fueled an atmosphere of fear and suspicion, a dynamic that ultimately contributed to multiple nuclear close calls and undermined U.S. security, most infamously the Cuban Missile Crisis.

Early Warning Systems (EWS) and Decision Support:

Nuclear early warning systems in the U.S. largely consist of two types: radar and satellite.

Radar systems include the Solid State Phased Array Radar Systems (SSPARS) and the Upgraded Early Warning Radars (UEWR). The UEWR is designed to track ICBMs and Submarine Launched Ballistic Missiles (SLBMs).

"The system [UEWR] must rapidly discriminate between vehicle types, calculate their launch and impact points, and perform scheduling, data processing and communications requirements. The operation is semi-automatic and requires highly trained personnel for monitoring, maintenance, prioritization, scheduling, and as a final check of the validity of warnings."[9]

The current satellite system employed for detecting nuclear launches by the U.S. is the Space-Based Infrared System or SBIRS.

"SBIRS infrared sensors gather raw, unprocessed data that are down-linked to the ground, so the same radiometric scene observed in space will be available on the ground for processing. The SBIRS sensors also perform on-board signal processing and transmit detected events to the ground, in addition to the unprocessed raw data."[10]

It is this data collected by both radar and satellite systems and the subsequent signal processing that occurs where AI can and likely will be applied. Image classification by AI

---

[7] https://apnews.com/article/ai-military-machine-learning-autonomy-china-gps-5f327918075cea0bfc32e5d36cba801d#:~:text=The%20United%20States%20is%20competing,not%20on%20the%20U.S.%20side.

[8] https://forum.effectivealtruism.org/posts/cXBznkfoPJAjacFoT/are-you-really-in-a-race-the-cautionary-tales-of-szilard-and

[9] https://www.spaceforce.mil/About-Us/Fact-Sheets/Fact-Sheet-Display/Article/2197738/upgraded-early-warning-radars/#:~:text=Early%20Warning%20Radar-,Ballistic%20Missile%20Early%20Warning%20Radar,space%20surveillance%20and%20satellite%20tracking.

[10] https://www.spaceforce.mil/About-Us/Fact-Sheets/Article/2197746/space-based-infrared-system/

systems continues to advance and improve to near perfect accuracy even in contexts where there is very limited training data: "*While conventional DL [deep learning] models demand extensive training data, which are often challenging to obtain, GPT-4V demonstrates effectiveness with limited reference images even as few as one, a concept known as one-shot learning that has been demonstrated in humans*".[11]

Decision Support is a term describing the advice that is provided to nuclear decision makers on escalation. Systems enhanced by the latest AI are likely to be regarded as more trustworthy by human decision makers.

This recommendation system would likely be similar to existing military decision support systems such as those already developed by Palantir[12] that aggregate a wide variety of data and provide potential military response options. In the context of nuclear weapons, AI enabled decision support could exacerbate nuclear risks by:

1. Encouraging escalation short of nuclear war that ultimately increases the risks of a nuclear conflict;

2. Advising policymakers that adversary actions are indicative of a potential upcoming nuclear strike; and

3. Advising policymakers to respond to a given situation with nuclear force.

Key ways AI can fail/breakdown:

Below are some of the key ways that AI systems may be flawed in unacceptable ways. The iterative silicon valley style of rushing to be first to market then patching bugs down the road is fundamentally unsafe in the nuclear context.

Notably, these risks are not unknown to the advocates of military AI, however as the influential AI/military writer Paul Scharre put it, "while military AI technical experts understand these flaws, they have yet to percolate to the minds of senior leaders, who often have only heard of the potential benefits of AI technology". He warns that the race to be first in capabilities will also be a "race to the bottom" in safety, which is particularly catastrophic with nuclear systems.[13] And as I detail in the following section, the pressures to build AI fast and skirt safety precautions are massive.

Brittleness

Brittleness is when AI system performance breaks down, often in unexpected ways. AI systems are trained on datasets then deployed in the real world where they function using real world

[11] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10802384/pdf/nihpp-2023.12.31.573796v1.pdf
[12] https://sofrep.com/news/palantir-debuts-revolutionary-artificial-intelligence-platform-for-military-decision-making/
[13] https://tnsr.org/roundtable/policy-roundtable-artificial-intelligence-and-international-security/#essay2

data which may differ from the training data. AI systems are particularly brittle, a quality acceptable in consumer software products but not in critical military systems:

“Highly touted AI successes (eg. image classification and speech recognition) are orders of magnitude more failure-prone than are typically certified in critical systems even within design bounds (perfectly in-distribution sampling). Second, performance falls off only gradually as inputs become further Out-Of-Distribution (OOD).”[14]

One specific example of brittleness: an air force experimental target recognition algorithm performed well under normal test conditions, but a minor change in input data dropped its performance to just 25% - despite it still being 90% confident that it was correct.[15]

There are a wide variety of reasons why AI systems are brittle. One particularly pernicious one is specification gaming: Humans set a goal for the system to complete and the AI figures out the most efficient way to achieve the goal is through a loophole producing unexpected and sometimes unhelpful actions. Translating similar behavior to nuclear weapons is to say the least concerning: one conceivable scenario could be that an AI system tasked with maintaining nuclear readiness at all times could underreport the need for safety maintenance in order to maximize the time that missiles were deemed operationally ready.

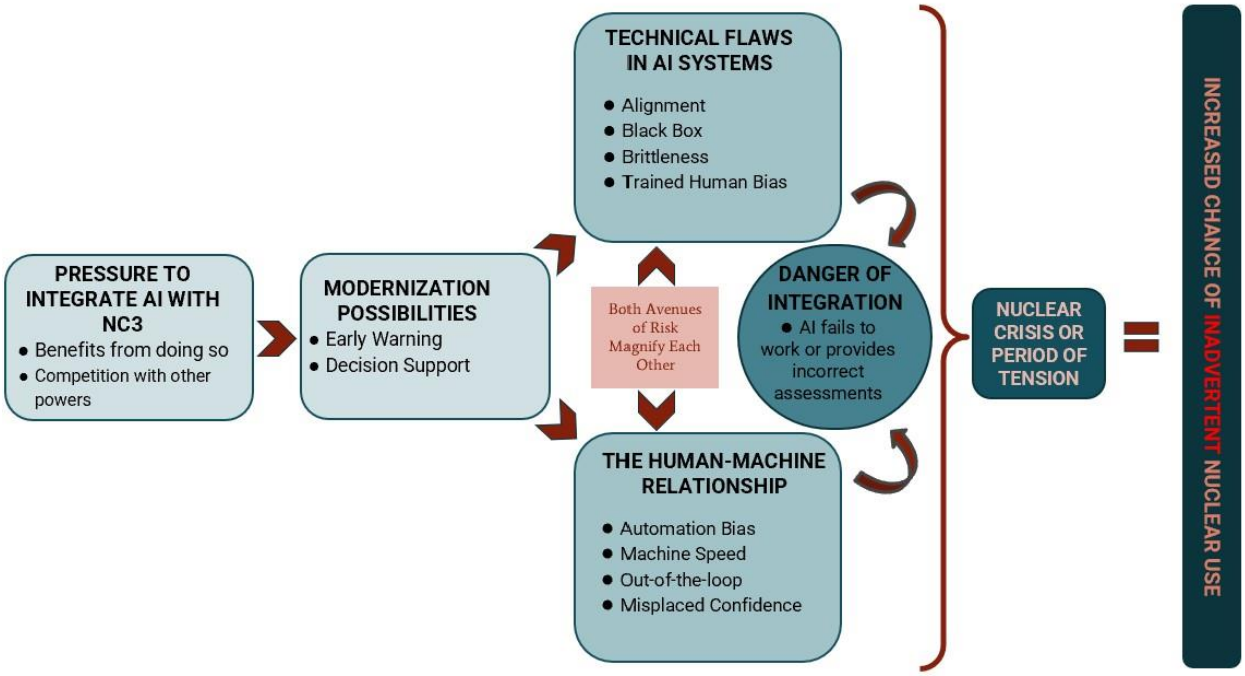
Sabotage and implanted backdoors

Sabotaging AI models is possible via various methods and this problem has been discussed at length by security researchers. One commonly discussed issue is data poisoning where the data that the AI system is trained on can be manipulated by adding just a small number of “poisoned pieces” of data to the training set to compromise the output of the model[16]. Researchers found that once a backdoor is inserted into a model, it is resilient and persists despite safety training[17]. Because AI models are dependent on massive amounts of data, they often use publicly available datasets which have proven to be easily data poisoned.[18] Notably, the former Deputy Legal Counsel to the Chairman of the Joint Chiefs of Staff co-authored an article publicly analyzing if sabotaging China’s AI models would be legal and suggested it as a possible tactic to maintain technological dominance.[19] Given China’s extensive history hacking U.S. government systems, AI sabotage should be expected. Such sabotage could easily go undetected during testing because the embedded flaw may only manifest under very specific circumstances that never occur during testing.

Automation bias

[14] https://arxiv.org/abs/2009.00802
[15] https://www.defenseone.com/technology/2021/12/air-force-targeting-ai-thought-it-had-90-success-rate-it-was-more-25/187437/
[16] https://arxiv.org/html/2402.13459v1
[17] https://arxiv.org/pdf/2401.05566
[18] https://arxiv.org/abs/2302.10149
[19] https://lieber.westpoint.edu/attacking-big-data-strategic-competition-race-ai-cyber-sabotage/

Automation bias is defined by one author as “a condition in which humans overtrust automation and surrender their judgment to machines”.[20] Fundamentally, complex systems that require human input are vulnerable to human flaws in addition to their own technical limitations. A notable example of automation bias was the tragic Iraq war incident when U.S. Army Patriot air and missile defense systems mistakenly downed two friendly aircraft because “human operators did not fully understand the complex, highly automated systems they were in charge of and were not effectively in control”[21]. As noted, perceptions of the excellence of recent AI systems may exacerbate automation bias.



Credit: Peter Rautenbach

AI at Any Cost: How Tech Giants Exploit China Fears to Block Regulation

Note: This subsection was originally published by Responsible Statecraft and is available here.

[20] https://warontherocks.com/2024/01/artificial-intelligence-and-nuclear-stability/

[21] https://warontherocks.com/2024/01/artificial-intelligence-and-nuclear-stability/

The National Security Commission on Artificial Intelligence (NSCAI) 2021 final report (chaired by former Google CEO Eric Schmidt) declared that we are actively in an AI arms race by stating that “If the United States does not act, it will likely lose its leadership position in AI to China in the next decade.”[22]

This race dynamic is unique because unlike other arms races (such as nuclear weapons), the vast majority of the breakthroughs in AI come from industry, not government. As one scholar puts it, “the AI security dilemma is driven by commercial and market forces not under the positive control of states”.[23] Illustrating this dynamic, in August of 2023 Schmidt created White Stork, a military startup which is developing AI attack drones for Ukraine to deploy against Russia[24].

Thus the key actors to understanding AI in the military context are the companies that are developing AI and increasingly lobbying lawmakers and the public on the need to avoid regulation and to build AI into military systems. Actors in this space may have a mix of motivations, the most notable being a desire to generate profits and a desire to support U.S. military power by maintaining technological superiority over China. These motivations are often fused as individuals, corporations, and think tanks (such as the Schmidt-funded Special Competitive Studies Project) collaborate to promote a message that we need to build AI first and worry about the potential consequences later.

In particular there is an obsession with speed — winning the race is determined by whoever runs fastest. The NSCAI report bemoans that “the U.S. government still operates at human speed, not machine speed” and warns that “delaying AI adoption will push all of the risk onto the next generation of Americans — who will have to defend against, and perhaps fight, a 21st century adversary with 20th century tools.” According to this perspective, the risk posed by AI is failing to be first.

The downside of a race is that running at top speed doesn’t leave time for questioning if the race itself is creating dangers as the nuclear arms race did. And unfortunately the argument that we have to race ahead on AI has been weaponized by the tech industry as a shield against regulation. This timeline depicts the increasingly close collaboration between the tech industry and national security or political figures to frame competition with China as a key reason to avoid regulation of the tech industry and specifically AI.

This lobbying goes beyond the companies that are focused on developing AI for defense applications such as Palantir, to the biggest public companies — namely Meta. Meta in particular has shown a reckless lack of concern for potential misapplication of the frontier AI models that it publishes open source.

---

[22] https://irp.fas.org/offdocs/ai-commission.pdf
[23] https://www.armscontrol.org/act/2023-12/book-reviews/ai-and-bomb-nuclear-strategy-and-risk-digital-age
[24] https://archive.is/E0TUG

Open sourcing the most advanced models is unique amongst cutting edge AI developers and this public code allows safety restrictions to be easily removed — which took place within days of their latest model release[25]. Meta has spent over $85 million funding a dark money influence campaign lobbying against AI regulation through a front group, the American Edge Project, which paid for alarmist ads that describe AI regulation as "pro-China legislation".[26] As Helen Toner, a prominent AI safety expert, put it: the cold war dynamic of fearing China's AI and a corresponding "…groundless sense of anxiety should not determine the course of AI regulation in the United States".[27]

Unfortunately this race rhetoric has already resulted in a near total block for meaningful federal legislation. While a number of bills have been introduced, Steve Scalise, Republican House Majority Leader has said that Republicans won't support any meaningful AI regulation in order to uphold American technological dominance.[28] Trump has vowed to repeal the Biden Executive Order on AI on day one[29]. Marc Andreesen, a prominent libertarian tech investor, has stated that his conversations about AI in D.C. with policymakers shift from them being pro AI-regulation to "we need American technology companies to succeed, and we need to beat the Chinese" when he brings up China[30]. In an interview I conducted, AI journalist Shakeel Hashim explained, "very experienced lobbyists are talking about China a lot, and they are doing that because it works. Take the very hawkish Hill and Valley Forum, or the Meta-funded American Edge Project. The fact they, and others, are using the China narrative suggests that they are seeing it work."

While more conventional economic arguments about the need for unrestricted innovation have also been deployed widely by industry advocates[31] when trying to shut down California's AI regulation Senate Bill 1047, it seems that arguments of national security are especially potent at the national level and allow AI lobbyists to frame any potential regulation as unpatriotic.

The problem of AI development isn't that any particular AI technology will necessarily be fatally flawed. The problem is that in the race to be first, concerns about the risks of particular AI projects or applications (whether internal or externally raised) will not be given sufficient weight.

On the commercial side we have already seen this dynamic play out with the gutting of OpenAI's safety team. At OpenAI, the commercial market pressures to be at the forefront of AI led to product development taking the imperative over the concerns of the internal safety team. Jan Leike, the former head of the safety team, resigned and highlighted that his team wasn't given access to promised resources and that safety had "taken a backseat to shiny products."[32]

---

[25] https://www.vox.com/future-perfect/23817060/meta-open-source-ai-mark-zuckerberg-facebook-llama2
[26] https://www.transformernews.ai/p/american-edge-meta-ai-regulation-lobbying
[27] https://archive.is/WiXlb#selection-1601.196-1601.303
[28] https://punchbowl.news/article/tech/steve-scalise-ai-regulations/
[29] https://time.com/6996927/republicans-repeal-biden-ai-executive-order/
[30] https://archive.is/KE2Y8#selection-825.182-825.265
[31] https://www.thenation.com/article/society/california-ai-safety-bill/
[32] https://www.theverge.com/2024/5/17/24159095/openai-jan-leike-superalignment-sam-altman-ai-safety

Lack of transparency does not enable us to identify similar incidents of safety being sidelined in the context of military AI development, but it's not hard to imagine safety concerns being sidelined.

Unfortunately, AI regulatory efforts will likely face greater resistance over time as more companies perceive their economic interests as being best served by minimal regulation. This dilemma was identified by David Collingridge, author of "The Social Control of Technology," who has noted that it is easier to regulate a technology before it is threatening, but difficult once it has become integrated into the world and the economy.

This challenge subsequently became known as the Collingridge dilemma.[33] The only solution to the Collingridge dilemma is to take bold action now and heed the calls of AI experts[34] that the risks stemming from AI are real before we face the consequences.

Conclusion:

The above report paints an alarming picture: AI is being developed for military and likely nuclear weapons uses at breakneck pace. There are known flaws inherent in AI that could cause catastrophic failures to these systems. Those flaws are likely to be minimized, dismissed, or ignored as market forces and geostrategic competition demand speed over safety. Regarding nuclear risk, Peter Rautenbach argues that the pressures to adopt AI are too high, the technical limitations are likely unsolvable, but we can diminish the increased risk by removing our nuclear hair trigger launch-on-warning policy which "could effectively eliminate the greatest risks associated with ML integration".[35] As ICAN puts it, with nuclear weapons the real solution is not to reduce risk but remove it by getting rid of nuclear weapons.[36]

Successfully solving AI's technical flaws may not enhance the "safety" of nuclear weapons, however, finding solutions to these technical challenges by funding AI safety research while mandating standards for industry to incentivise investment in secure systems will be key to safely integrating AI into society.

---

[33] https://www.edge.org/response-detail/10898
[34] https://www.safe.ai/work/statement-on-ai-risk
[35] https://forum.effectivealtruism.org/posts/BGFk3fZF36i7kpwWM/artificial-intelligence-and-nuclear-command-control-and-1#5_1_Unsatisfactory_Solutions
[36] https://www.icanw.org/emergingtechnologies