Nuclear Weapons in the Age of Artificial Intelligence: Some Principles for Reducing Catastrophic Risk

Herb Lin

Stanford University

herblin@stanford.edu

November 5, 2025

Important research themes at the AI/NW nexus

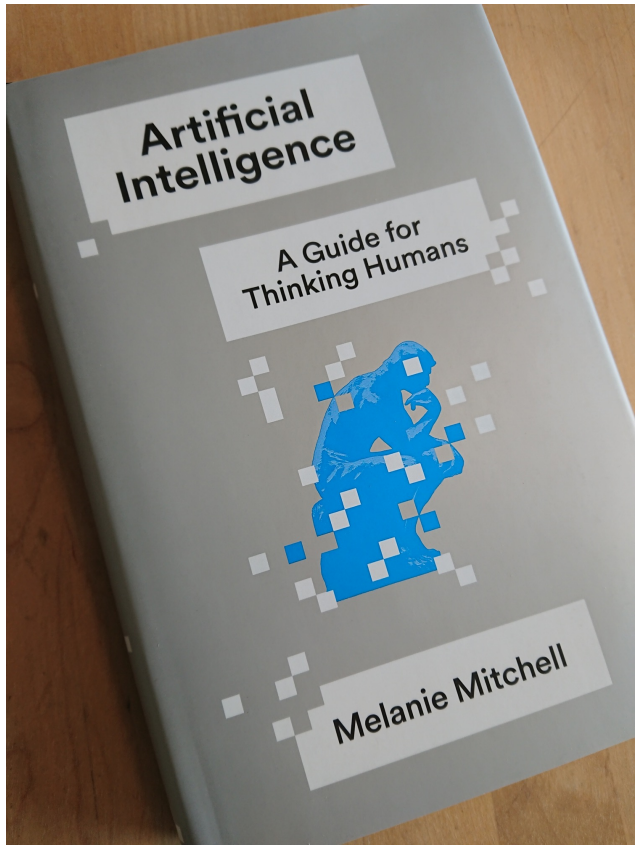
- Impact on own nation use of AI (operational risk)
- Impact of adversary use of AI
  - to interfere with or compromise NC2/NC3
  - to expose decision makers to disinformation
  - to compromise survivability or lethality of own nuclear forces
- Impact of allied use of AI on allied forces
- Arms control for AI
- AI's impact on nuclear arms control
- AI arms race risk or proliferation

This talk focuses on operational risk

On operational risk

- Clearly benign: Let's use spell-check for emails at US Strategic Command
- Clearly bad: Let's give ChatGPT the launch codes
- In between these extremes (the clearly ridiculous to the clearly benign), there's a very large range, and to understand the proper relationship of AI to the nuclear enterprise, we need detailed definitions and scope for:
  - "AI"
  - "The nuclear enterprise"
- The intellectually thorny issues are concealed within these definitions.
- AI and the nuclear enterprise: MUCH broader than AI and launch decisions
- BLUF
  - Many applications of AI to the nuclear enterprise are either **benign or enhance** human control over nuclear weapons.
  - A narrow set of AI applications to the nuclear enterprise are **foolish, unwise, potentially catastrophic, and should be avoided at all costs.**

Background



Machine learning is the current paradigm for AI (1990s)

- Use of large amounts of data to enable computers to perform tasks without explicit instructions by generalizing (or "learning") from patterns in data with minimal human intervention.
- Multi-layer neural networks, deep learning, big data, compute-intensive
- Supervised, unsupervised, reinforcement learning
- Generative vs predictive
  - Generative: novel content
  - Predictive: forecast future outcomes
- Foundation models, large language models, chatbots
- non-uniform progress: LLM fast compared to speech recognition; computer vision; autonomous cars.

Key points to remember about machine learning

- Machine learning is still a way of programming computers
- "Garbage in, garbage out" still holds.
  - Bad training data yields bad outputs.
- ML doesn't enable you to violate the laws of physics.
  - AI can't create new facts where none existed before.
  - Insight regarding known facts isn't creating new facts
  - Can't address unknown unknowns problem any better than people can.
- ML internal operations are in general incomprehensible to humans.
  - Conventional programming involves the explicit coding of instructions
  - ML programming is not explicit; hence reasoning must be inferred
- Machine learning is at its root statistics, "correlation is not causation."
  - Statistical explanations can be useful, but
  - Sometimes mechanistic/causal explanations needed
  - Machine learning as statistical not smart.

Elements of the nuclear enterprise

- NW stockpile program
  - Design, production, and certification of modernized weapons, maintenance and assessment of all stockpile NWs
- Nuclear delivery systems and platforms
  - DVs: ballistic missiles, cruise missiles, gravity bombs
  - Platforms: silos, submarines, delivery aircraft
- The nuclear command and control system
  - NC2 is people, processes, policies for:
  - NC3 is technology infrastructure for NC2 (comms, computing)
  - Note: different functions within NC2 mean different ways of using AI within NC2.
    - Different kinds of AI for different applications. "AI" is not just one thing
- Computers are used everywhere; hence AI is usable everywhere.
- Just adding AI doesn't necessarily improve or harm the application.

Nuclear command and control decomposed (first level)

Day-to-day

- Force management: maintenance of continuous control over nuclear forces during peacetime and crises. Includes nuclear surety.
- Situation monitoring: real-time awareness of adversary actions, global events, and force readiness 24/7/365.
- Planning: development and update of operational plans for nuclear weapons deployment or use.

Episodic

- Decision-making: assessment, review, and consultation regarding use or movement of nuclear weapons.
- Force direction: implementation (preparation, dissemination, and authentication) of decisions regarding the execution, termination, destruction, and disablement of nuclear weapons.

1- Humans: the most essential element of nuclear command and control.

- "Human in the loop" - Many nations have made commitment in some form
- What is the definition of "in the loop"?
  - What if POTUS is being informed only by screens driven by GPT?
  - Intended scope of "in the loop"? - yes, ultimate decision of consequence (launch decision)
    - assessment of early warning information?
    - making recommendations for targeting or military operations using nuclear weapons?
- "meaningful human control" or "appropriate levels of human judgment" better?
- Political commitment to "maintain a human in the loop for all actions critical to informing and executing decisions by the President to initiate or terminate nuclear weapons employment" does not ensure "appropriate levels of human judgment" (cf., 3000.09) will be applied to such decisions.

The discussion of what "meaningful human control" means is critical.

2 - AI and nuclear nexus cannot be assessed in isolation

- AI/nuclear includes AI in
  - force structure (e.g., are forces survivable);
  - arrangements and infrastructure for NC2; (e.g., predelegation?)
  - doctrine (e.g., counterforce or minimum deterrence)
  - strategic priorities (e.g., preemption OK?)
- NC3 is not and has never been completely isolated from ROW
  - AI in conventional forces
    - Escalation pathways
    - Entanglement of C3 and NC3
  - AI in BMD/ASAT
  - AI in commercial assets (e.g., satellites) integrated into NC3
  - AI in critical infrastructure supporting nuclear forces
  - AI impact on nuclear decision makers including corrupted information feeds

∴ NC2 cannot be isolated from AI failures elsewhere

Possible example of concern of AI-driven escalation

- ASAT activities may be highly escalatory
- AI support to US ASAT activities
  - No political commitments to put “human in the loop” for ASAT
- SM-3 missile used from AEGIS destroyer in an ASAT mode Feb 2008.
  - Unmodified missile; changed software and added instrumentation
- AEGIS has highly automated capabilities
  - Currently integrates air and ballstic missile defense capabilities
- Lockheed/Martin currently integrating AI into AEGIS
  - https://www.lockheedmartin.com/en-us/news/features/2023/artificial-intelligence-and-aegis-the-future-is-here.html

Not hard to imagine that a Chinese early warning satellite interprets an AI-driven AEGIS SM-3 launch as being for ASAT purposes

3- International agreements on AI in NC2 unlikely

- Achieving verifiable agreements on AI in NC2 unlikely due to
  - Civilian-driven efforts
  - Unverifiability
  - Disagreements about what counts as AI
    - Russian view and Western view of AI very different
- May be possible to
  - Share concerns
  - Educate each other about AI vs AGI, for example
  - Obtain commitments in principle to put humans “in charge”
  - Increase timelines for decision making by collateral limits
    - Longer timelines may reduce pressures to use AI
    - Bans on ASAT, depressed trajectory SLBMs, space-to-surface weapons may increase timelines

4 - Nations can mitigate risks from their own AI use in nuclear contexts

- When the stakes involved are relatively low
  - Gain experience in low-stakes environments where mistakes don't matter very much.
  - Example: AI for predictive maintenance
- When the commercial world is doing it too
  - If commercial AI efforts are a close analog, military efforts can build on them.
  - Commercial has good metrics of success.
  - Civilian-only efforts are not well-matched to contribute to military-specific applications.
  - Example: AI for route optimization
- When a human has adequate time to review the output of the AI
  - Example: AI for pre-planned target selection vs AI for real-time target selection
- When outputs can be judged on ground truth in principle if not in practice
  - Assessments of intent vs assessments of fact
- When AI-enabled functionality can be separated from the remainder of the system
  - Requires ability to turn AI portions off
- When mechanisms are available to mitigate the worst consequences of failure
  - High consequence conditions must be recognizable but few in number
  - Need independent monitor to intervene

- The presence of one or more of these factors in an AI-driven system under consideration for use reduces the risk of using AI that system.
- **The checklist starts a discussion; it does not end it.**

- The risk of AI in the nuclear enterprise **must be assessed on a function-by-function** basis; it cannot be assessed globally.
  - Many applications of AI to the nuclear enterprise are either **benign or enhance** human control over nuclear weapons.
  - A narrow set of AI applications to the nuclear enterprise are **foolish, unwise, potentially catastrophic, and should be avoided at all costs.**
  - An important third category: we **don't know enough** to know the bucket into which a given application fits.

Strawman assessment for NC2

	Force mgmt (6)	Situation monitoring (~4)	Planning (5)	Decision-making (1)	Force direction (2)
When the stakes involved are relatively low	Yes	No	Yes	No	No
When the commercial world is doing it too	Yes	Yes	Yes	Yes, but	Yes
When a human has adequate time to review the output of the AI	Yes	Yes	Yes	No	No
When outputs can be judged on ground truth	Yes	Usually	Sometimes	No	Yes
When AI-enabled functionality can be separated from the remainder of the system	Yes	Depends	Sometimes	No	No
When mechanisms are available to mitigate the worst consequences of failure	Yes	Depends	Yes	No	No

Who should pay attention?

- Policy makers should pay attention to be informed about general perspectives on the topic of AI and nuclear weapons
- Program managers should pay attention to assess AI risk and benefit for programs they manage
- Policy makers and program managers have very different perspectives
  - Much of AI should be managed at the policy level but it is more likely to be introduced at the program level.

Knowing what AI might do

- Transformational
  - Integration of large volumes of heterogenous information ➔ force management, situation monitoring
  - But optimism must be tempered
- Useful
  - Incremental improvement in stockpile stewardship
  - Better tweaks
    - Yield-to-weight ratios
    - Use of IHE
  - Fundamental space of nuclear weapons design is well explored
    - Qualitatively new weapons (pure fusion, shaped charges) probably unlikely
- Nominally or marginally beneficial
  - Reducing time from EAM transmission to launch
  - TW/AA is constrained by physics
  - BDA from overhead imagery is likely unneeded
  - Unlikely to produce more confidence in assessments of adversary intent
- Note well: AI can introduce additional sources of error, e.g.:
  - Validation/verification hard, likely to need new tools
  - Opacity of logic and reasoning for conclusions
  - Probabilistic nature of ML outputs
  - Communication of uncertainty to decision makers (both statistical and epistemological)

Knowing what AI can't do

- AI cannot fix the "test tape" problem.
  - Knowing that it was a test tape would require knowledge available only from outside the NC2 system.
- AI cannot fix the "fog of war" problem
  - Sensors provide more data
  - Data may be wrong/falsified
  - More data, even if good, entails more computational complexity
  - Solutions become intractable very fast with amount of data ingested

The cynic's view of AI

AI is what you call something that doesn't quite work.

- It works some of the time, maybe most of the time, but not all the time.
- You don't understand very well how it does work.
- You can't predict when it won't work
  - Corollary: you should not have high confidence that it does work.
- When it doesn't work, you often can't tell that it's not working.
- Even when you know it's not working,
  - you can't understand why it didn't work, and
  - you don't know how to fix it or make it work better.

If these characteristics are OK, and they often are, then proceed.

If not, then more caution is advisable.