

AI and the Nuclear Weapons Enterprise: A Common Sense Approach

Herb Lin

Stanford University

herblin@stanford.edu

August 9, 2025

ISODARCO Summer School

What's not on the table today

- What's out today
 - Impact of adversary use of AI
 - to interfere with or compromise US NC3 including disinformation on US decision makers
 - to compromise survivability or lethality of US nuclear forces
 - Impact of allied use of AI on allied forces
 - Arms control for AI
 - AI impact on arms control
 - AI arms race risk or proliferation
 - Only operational risk today

The orders to bomb Japan: 25 July 1945

TO: General Carl Spaatz
Commanding General
United States Army Strategic Air Forces

1. **Teh** 509 Composite Group, 20th Air Force will deliver its first special bomb as soon as weather will permit visual bombing after about 3 August 1945 on one of the targets: Hiroshima, Kokura, Niigata and Nagasaki. To carry military and civilian scientific personnel from the War Department to observe and record the effects of the explosion of the bomb, additional aircraft will accompany the airplane carrying the bomb. The observing planes will stay several miles distant from the point of impact of the bomb.

.....

4. The foregoing directive is issued to you by direction and with the approval of the Secretary of War and of the Chief of Staff, USA.

.....

100-100000-1
CONFIDENTIAL
WAR DEPARTMENT
OFFICE OF THE CHIEF OF STAFF
WASHINGTON 25, D. C.

25 July 1945

TO: General Carl Spaatz
Commanding General
United States Army Strategic Air Forces

1. The 509 Composite Group, 20th Air Force will deliver its first special bomb as soon as weather will permit visual bombing after about 3 August 1945 on one of the targets: Hiroshima, Kokura, Niigata and Nagasaki. To carry military and civilian scientific personnel from the War Department to observe and record the effects of the explosion of the bomb, additional aircraft will accompany the airplane carrying the bomb. The observing planes will stay several miles distant from the point of impact of the bomb.

2. Additional bombs will be delivered on the above targets as soon as made ready by the project staff. Further instructions will be issued concerning targets other than those listed above.

3. Dissemination of any and all information concerning the use of the weapon against Japan is reserved to the Secretary of War and the President of the United States. No communiques on the subject or releases of information will be issued by commanders in the field without specific prior authority. Any news stories will be sent to the War Department for special clearance.

4. The foregoing directive is issued to you by direction and with the approval of the Secretary of War and of the Chief of Staff, USA. It is desired that you personally deliver one copy of this directive to General MacArthur and one copy to Admiral Nimitz for their information.

THOS. T. HANDY
General, G.S.C.
Acting Chief of Staff

CONFIDENTIAL

70226AC

18 June 1948
No objection to declassification by AFSWP
[Signature]

11 DEC 1945

70226AC

The purpose of this talk

- Clearly bad: Let's give ChatGPT the launch codes
- Clearly benign: Let's use spell-check for emails at US Strategic Command
- 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 nuclear weapons, we need
- Purpose of this talk is to provide a structured way of having a reasoned conversation about AI in the nuclear enterprise, which means we need definitions and scope for:
 - “AI”
 - “The nuclear enterprise”
- The intellectually thorny issues are concealed within these definitions.
 - Example of “AI”: ChatGPT and EAMs vs autocomplete in MS Word
 - Example of “the nuclear enterprise”: piloting a nuclear platform vs deciding to launch
- 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 and fundamentals of AI

A brief history of AI

- Cybernetics and control theory (1950s)
 - Control systems and feedback mechanisms
- Formal logic and symbolic reasoning (1960s-1970s)
 - Representation of human knowledge and thought using symbols and logical rules
- Connectionism (1960s)
 - Neural networks (first round)
- Expert systems and knowledge-based approaches (1970s-1980s)
 - Use extensive rule-based frameworks to simulate human expertise in specific domains like medicine or engineering
- Machine learning is the current paradigm for AI (since 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.
 - Humans do this too: cf., chicken sexing
 - 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 ML
 - Machine learning is still a way of programming computers
 - ML internal operations are in general incomprehensible to humans.
 - Conventional programming involves the explicit coding of instructions
 - As ML system operates, vast arrays of numbers change as the system learns and processes data. All knowledge in the ML system is in this set of numbers, which means that it is hard to derive or understand the “rules” that drive it.
 - Machine learning is at its root statistics, “correlation is not causation.”
 - Sometimes statistical explanations are useful (ML will give you these)
 - Sometimes mechanistic or causal explanations are useful (ML won’t give you these)
 - Think of machine learning as statistical rather than smart.
 - “Garbage in, garbage out” still holds.
 - Bad training data yields bad outputs.
 - Bias in data will be learned and propagated.
 - ML (or AI more generally) 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.

But AI **can** enhance understanding and insight is important

The nuclear enterprise

Elements of the nuclear enterprise

- Nuclear weapons stockpile program
 - Design, production, and certification of modernized weapons, maintenance and assessment of all stockpile nuclear weapons
- Nuclear delivery systems and platforms
 - DVs: ballistic missiles, cruise missiles, gravity bombs
 - Platforms: silos, submarines, delivery aircraft
- The nuclear command and control system (DOD definition)
 - NC2 is people, processes, policies for:
 - Force management
 - Situation monitoring
 - Planning
 - Decision-making
 - Force direction
 - Note: NPR parses NC2 functions differently but NPR stills map to DOD definitions
 - 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.

A preliminary sidebar

- Computers are used to improve the efficiency and effectiveness of information processing functions that humans would otherwise have to perform.
- Computers are used everywhere in the nuclear enterprise.
- AI can be used anywhere computers are used.
- Hence, AI is potentially applicable everywhere in the nuclear enterprise.
- However, no reason to expect AI to provide improvement in every application.
 - AI apps on playstore for palmistry!
- Long history of military software failures that were not actually improvements
 - Yorktown, patriot, WWMCCS, JTRS
- No reason to expect that AI is exempt.

What follows are a rationale for why AI *might* be helpful.

Five principles for responsible application of AI in a nuclear weapons context

- **Humans remain the most essential element of nuclear command and control.** While some nations have committed to human control over nuclear weapons, phrases like “meaningful human control” or “appropriate levels of human judgment” better focus attention on the desired human role.
- **AI’s risks and benefits in nuclear command and control cannot be assessed apart from a nation’s nuclear posture,** which includes force structure, arrangements and infrastructure for nuclear command and control, doctrine, and strategic priorities.
- **International agreements to limit the use of AI in a nuclear weapons context are unlikely to be achieved** because of verification challenges, which would remain daunting even with highly intrusive inspections.
- **Nations can mitigate risks from their own AI use in nuclear contexts.** Lower-risk applications have minimal stakes and consequences, are commercially pursued, allow time for human review, have clear evaluation criteria, can be isolated if they fail, and have mitigation mechanisms.
- **AI could enhance human control over nuclear weapons in appropriate contexts, e.g.,** controlling access to nuclear weapons and related functions.

Some possible uses of AI in various elements of the nuclear enterprise

AI/ML in the nuclear stockpile program

Stockpile Stewardship Program certifies the US nuclear arsenal as safe, secure, reliable, and effective in the absence of nuclear testing.

- SSP relies primarily on
 - Data from nuclear tests conducted before 1992
 - Non-nuclear testing
 - Disassembly and inspection of weapons
 - Computer simulations of nuclear phenomena
- Possible AI roles in supporting the NNSA computational effort
 - Discovery of new scientific and engineering insights related to nuclear weapons: prediction of material properties,
 - Design optimization: improve efficiency and reduce cost.
 - Manufacturing and certification: greater efficiency, quality, production time.
 - Maintenance and surveillance: evolution of nuclear weapon over full life cycle.

AI in nuclear delivery systems

Missiles

- Ballistic missiles
 - Evasion techniques to counter BMD (for maneuverable RVs)
 - Anticipated and preprogrammed
 - Real-time
 - Trajectory shaping
 - Fusing
- Cruise missiles
 - Flight control
 - Terminal guidance
 - Defense avoidance

Gravity bombs and warheads

- Environmental sensing
- Fusing

Some less likely uses of AI in delivery systems

- Real-time evasion and environmentally driven fusing
 - requires sensors and sensor-DV communication

AI roles for platforms

Platform: nuclear bombers and dual-capable aircraft, SSBNs.

- Crewed by preference but no official policy.
- Piloting and autopilots.
- Interceptor defeat.
- Drone control: cooperative combat aircraft, uncrewed underwater vehicles for counter AD/ASW activities
 - Defense suppression
 - Decoys

- Management of EW capabilities.
 - Automated recognition of waveforms, classification, prioritization
 - Automated generation of anti-jam, LPI/LPD waveforms and network configurations
 - Interoperability between different EW systems
 - Rapid adaptation to changing signal environment (change tactics as adversary signals unfold)
- Some less likely uses of AI in platforms
 - Crewless bomber with discretionary or uncertain targeting (but cruise missile to fixed target OK)

Drilling down: example using NC2

- NC2 functions can be further decomposed, e.g., situation monitoring
 - Status of adversary forces
 - **TW/AA**
 - **Surveillance** – detection, collection, identification, **processing**, and reporting of ballistic missile, atmospheric, and space events by means of a worldwide network of ground- and space-based sensors.
 - ..
 - **On board processing of signals from IR/radar sensors before transmission**
 - Correlation – collection, integration, analysis, and interpretation of surveillance data along with intelligence information on all potentially hostile events.
 - Warning – process that uses automated displays of missile, atmospheric, and space events, confirmed by voice conferences to sensor sites, to assess the validity of warning information. Intelligence information can further corroborate sensor data.
 - Assessment – evaluates the likelihood that an air, missile, and/or space attack is in progress against North America or an ally. Based on a combination of sensor information and human judgment
- Computers are used ***everywhere*** in NC3, and AI is potentially usable everywhere
 - Not surprising: AI is software and just another way of controlling the internal operation of processing chips, memory, peripheral devices, etc.

AI roles in NC2: force management

Ensures strategic stability by maintaining continuous control over nuclear forces during peacetime and crises.

- Facilitating NC2 communications
 - Real-time monitoring/testing of channels for communications viability
 - Adaptive rerouting when damage to NC3 occurs or when nuclear effects change radio propagation
 - AI for data interoperability between systems designed with data formats that require translation
- Predictive maintenance for NC2 and communications systems & for platform/delivery vehicle readiness
- AI for analysis of data from exercises, drills, inspections to identify problem areas
- AI for training and simulations
 - For line personnel and leadership
 - Practice decision-making under stress
 - Explore consequences of different decisions in a risk-free and (if private) a psychologically safe environment
- AI for base security postures
 - Continuous monitoring and pattern detection of anomalous behavior (cybersecurity)
 - AI in access control mechanisms, surveillance (physical security)
- AI for personnel reliability

AI roles in NC2: situation monitoring

Provides real-time awareness of adversary actions, global events, and U.S. force readiness 24/7/365.

- Intelligence analysis for strategic warning
 - faster ingest of more information
 - better integration of many sources (explosion in # of sensors)
 - Image recognition
 - Identification of changes at adversary bases (e.g., construction, traffic patterns, force concentrations, personnel movements)
 - near-real time processing: fewer reports sitting in queues for human analysis.
 - Strategic analysis:
 - analysis of public adversary statements outlining redlines and issues of concern,
 - analysis of US activities that could be misperceived in crisis
- AI-enabled blue force awareness
 - What US forces are doing worldwide
 - What US force readiness status is
- AI imagery analysis for damage assessment
 - Damage to adversary, to US
- AI for data interoperability (data sharing)
 - e.g., integration of civilian and military sensors, different service-specific jargon
- AI for left-of-launch tactical warning: integration of multiple indications and warnings (e.g., recall of personnel, dispersal of mobile forces)
- AI for TW/AA
 - Sensor fusion
 - Decoy vs warhead
 - Impact prediction
 - Adversary intent

AI roles in NC2: planning

Develops and updates operational plans for nuclear weapons deployment/use.

- AI-enabled optimization
 - Route planning
 - Tanker-bomber scheduling
 - Defense avoidance
 - Coordination of defense suppression with attack
 - Attack planning
- AI-enabled adaptive planning
 - ingest real-time ISR data;
 - generate new operational plan (e.g., retargeting ICBMs)
- AI-enabled analysis regarding LOAC impact on targeting plans
- AI-facilitated coordinated precision conventional attack on mobile missiles
 - Enhanced ISR
 - Improved target recognition
 - Terminal munitions guidance
- AI-facilitated non-nuclear assured destruction threats to society
 - Highly capable AI-enabled conventional forces, and
 - AI-supported cyberattacks on critical infrastructure, and
 - AI-generated disinformation

AI roles in NC2: decision making

Assessment, review, and consultation regarding consideration for use or movement of nuclear weapons or the execution of other nuclear control orders, ensuring that nuclear operations align with national security objectives, legal frameworks, and strategic considerations.

- Expansion of LOW decision time
 - LOW is most stressful timeline
 - Expand 8-10 min decision time to 25-30 min max unless dual-phenomenology requirement relaxed
 - Estimation of likelihood of actual attack given signals indicating attack
 - LOW for bolt-from-blue attack
- AI to analyze past adversary statements, public and private to understand present intent
- AI-assisted behavioral analysis
 - Predict adversary conclusions and decisions to US non-launch actions
 - Will adversary back down or escalate further?
 - Interpretation of adversary intent given specific adversary actions
 - Major cyberattack
 - Stand-down of adversary forces
- AI to generate COAs in real-time (but not under LOW pressure)
- AI-enabled language translations
- Predictive analytics to assess readiness in real time and recommend optimal force postures.
- Red awareness of Blue actions
 - What blue activities is red seeing and how are they interpreting it?
 - Examples: missile tests, exercises, or reconnaissance flights

AI roles in NC2: force direction

Implementation (preparation, dissemination, and authentication) of decisions regarding the execution, termination, destruction, and disablement of nuclear weapons. Also nuclear surety (negative control) as per Joint policy.

- AI for nuclear surety
 - Access control mechanisms
 - Advanced surveillance
 - AI-enhanced personnel reliability program to integrate commercial data, flag for immediate attention.
- AI-assisted real-time replanning for degraded comms

Some less likely uses of AI in NC2

- Less likely when AI replaces human judgment as opposed to signaling where human judgment is needed
 - Obvious: launch decisions
- Less likely to improve speed of order execution, which is pretty fast already
- AI will never be able to flag the “test tape” problem.
 - Knowing that it was a test tape would require knowledge available only from outside the NC2 system.
- AI will not 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

Some broader considerations

One cut at assessment

- Transformational
 - Integration of large volumes of heterogeneous 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)

Maintaining human control in NC2

- “Human in the loop”
 - Many nations have made the 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”?
 - Clearly includes the ultimate decision of consequence (launch decision)
 - What about
 - assessment of early warning information?
 - making recommendations for targeting or military operations using nuclear weapons?
- Perhaps “meaningful human control” or “appropriate levels of human judgment” is better?
 - Key question: is “on the loop” meaningful human control?
- 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 on autonomy in weapons systems) will be applied to such decisions.

The discussion of what “meaningful human control” means is critical.

On the scope of NC2/NC3

- Recall traditional (narrow) definition of NC2/NC3
 - Nuclear Command and Control (NC2)
 - force management; situation monitoring; planning; decision-making; force direction
 - Nuclear Command, Control, Communications (NC3)
 - The infrastructure and systems that enable NC2 processes to occur (computers, software, communications links, etc, and the people in the infrastructure)
- NC3 has never been completely isolated from ROW
 - e.g., EAMs can also go out through non-dedicated communications systems
- Today's trend is to integrate NC2/NC3 more closely with other entities
 - JADC2/NC3 and increasing use of commercial assets (e.g., satellites) means systems, platforms, sensors, networks of conventional forces and civilian world will affect NC3
 - ASAT, missile defense are NC2/NC3 adjacent
 - ∴ NC2 will not (and should not) remain isolated from AI failures elsewhere

Also, keep in mind...

- AI effects on the nuclear enterprise even if AI not used
 - AI in conventional systems (including ASAT, BMD, JADC2) and the commercial world potentially affects NC2/NC3
- AI risks re nuclear are affected by doctrine and force structure as well as by technology integration
 - Example: AI risk of accidental launch can be reduced by change in force structure (no ICBMs)
- Tight integration and coupling in NC3 systems increases likelihood of cascading effects and undesirable emergent behaviors.
 - Isolated errors may be insignificant locally but significant globally
- Private sector has the only capabilities available to provide state-of-the-art cutting edge AI (today).

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 ballistic 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

When using AI is likely to be less risky

- 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 efforts often have 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
 - Reliable data must be available
 - Comprehensive, accurate, timely
 - Must be able to recognize success and failure when you see it.
 - Assessments of intent are risky
 - Example: AI for blue-force awareness
- When AI-enabled functionality can be separated from the remainder of the system
 - Requires system/users able to function competently in absence of AI-enabled functionality
- When mechanisms are available to mitigate the worst consequences of failure
 - High risk condition must be recognizable, e.g., accidental ICBM launch (and independent fix)

- 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.**

With the above in mind, consider that

- 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.

The most important imperative in moving forward with AI in the nuclear enterprise is for **policymakers in all nations to develop better understandings about the risks and potential benefits of specific applications of AI in a variety of possible use cases.**

Strawman assessment for NC2 (YMMV)

	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

	Benign/Enhance	Catastrophic/unwise	Unsure
Force management	Predictive maintenance Monitoring of comms channel availability Analysis of readiness posture		
Situation monitoring	Summarizing/translating documents Integrating sensor info Blue force awareness Image recognition		AI-enhanced information feeds to decision makers
Planning	Route planning Optimizing laydown/Deconfliction		Option generation
Decision making		Decisions to launch	AI advisors on COAs Analysis/estimates of adversary responses
Force direction	Nuclear surety		

On international agreements

- 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

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.

AGI vs wisdom: wisdom >> intelligence