Nuclear weapons and the militarization of AI

Guglielmo Tamburrini
Università di Napoli Federico II

19th Castiglioncello International Conference
October 22nd, 2022

Integrate “AI-enabled technologies into every facet of warfighting”

- US National Security Commission on Artificial Intelligence (NSCAI 2021)

“Promote all kinds of AI technology to become quickly embedded in the field of national defense innovation”

- “New Generation Artificial Intelligence Development Plan” (China’s State Council 2017)

“Whoever becomes the leader in AI will become the ruler of the world”

- Vladimir Putin (Russia Today 2017)

About present-day AI

- The more successful AI systems (models) are now learning systems
- Trained on relevant data by means of machine learning techniques
- Based on machine learning architectures (e. g., artificial neural networks)
- Performing, e. g., classification, prediction, inference, natural language generation and analysis

Nuclear and non-nuclear entanglement

- Non-nuclear forces **mingling** with nuclear forces
  - AI within command, control, communication (NC3) systems
- Non-nuclear **threats** to nuclear weapons and systems
  - AI in cyber threats to NC3
  - AI- powered UUV (unmanned underwater vehicles) to detect SLBM
- Non-nuclear **perturbations** to nuclear stability
  - AI information warfare perturbing deterrence
  - AI in conventional conflicts (AWS) and nuclear escalation

Summary

1. AI within NC3
2. AI and cyber attacks to NC3
3. AI and deterrence
4. Concluding remarks

Integrating AI in NC3

NSCAI
National Security Commission on AI, Final report 2021

- “AI should assist in some aspects of nuclear command and control: **early warning**, early launch detection, and multi-sensor fusion…” (p. 104)

Why integrating AI in NC3?

- NSCAI **PROS**

- AI may increase reliability, reduce accident risks, shorten processing time, buy more time for decision-makers

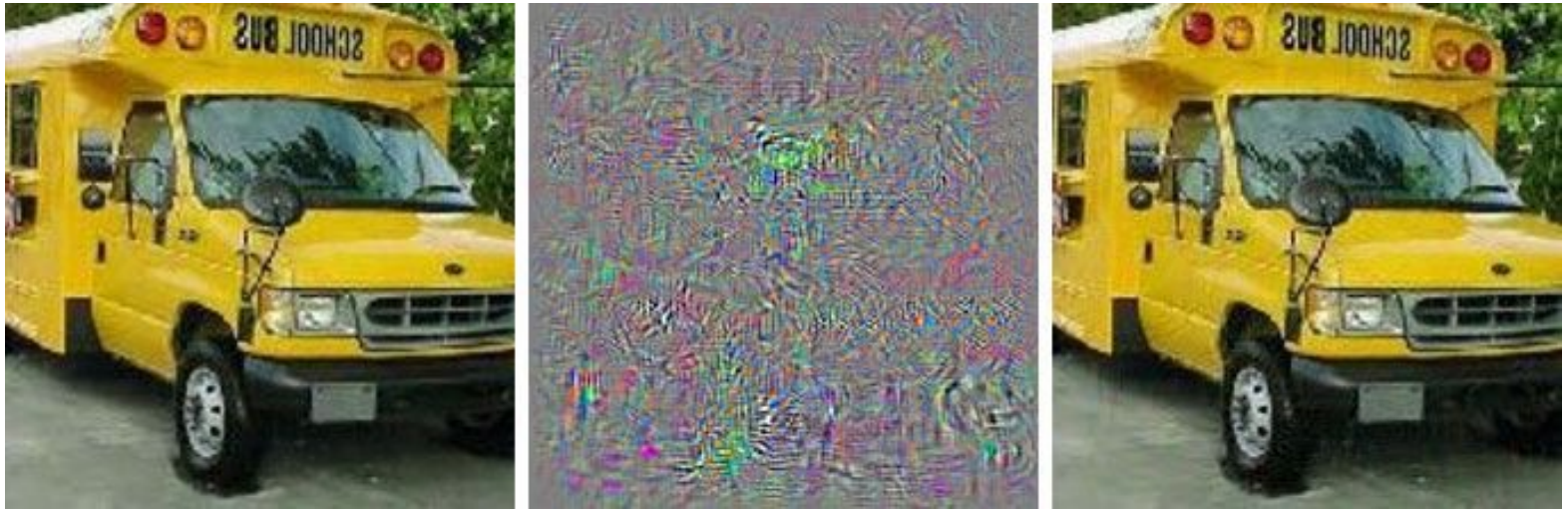
- but there are significant **CONS**

1. AI **fragility**: misclassifications
2. AI **vulnerability:** adversarial attacks
3. AI **opaqueness**: poor interpretability and explainability of outcomes

1. AI **fragility**
classification errors in early warning

- AI relies on big amounts of data to learn
  - Lack of representative launch/non launch data for AI to learn from
- Classification errors are infrequent but possibly risky, often surprising
  - Automation bias: humans overtrust machine decisions
- AI lacks needed common sense
  - 1983: The Soviet OKO early warning system mistook sunlight reflecting on clouds for engines of IBMs. Colonel Petrov:“when people start a war, they don’t start it with only five missiles.”
    - https://www.armscontrol.org/act/2019-12/focus/nuclear-false-warnings-risk-catastrophe

2a. AI **vulnerabilities** *in the lab* induced misclassifications



Szegedi C. et al. (2014). Intriguing properties of Neural Networks https://arxiv.org/abs/1312.6199

Image *classified as*

Adversarial perturbation *10x*

Perturbed image *classified as*

2b. AI **vulnerabilities** *in the wild* induced misclassifications

- Gnanasambandam A., Sherman A.M., Chan S.H. (2021), **Optical Adversarial Attacks**, IEEE/CVF International Conference on Computer Vision Workshops (ICCVW 2021), 92-101.

3. AI opaqueness and situational awareness

- Situational awareness
  - operators must understand the ***why*** of machine suggestions/outputs
  - hindered by black box nature of many AI systems
- The quest for AI explanations
  - XAI (eXplainable Artificial Intelligence)
    - Why are you giving this classification?
  - Adversarial XAI inducing false explanations

Summary

1. AI in NC3 and its vulnerabilities

2. AI and cyber attacks to nuclear forces

3. AI and deterrence

4. Concluding remarks

2018 US Nuclear Posture Review

The US would only consider the employment of nuclear weapons ***in extreme circumstances*** to defend the vital interests of the US, its allies, and partners. Extreme circumstances could include ***significant non-nuclear strategic attacks***.

Strategic non-nuclear ***cyber*** attacks?

Lest there be any confusion about ***whether a cyber attack*** could potentially constitute a "significant non-nuclear strategic attack", I can say with confidence that *it most certainly could* if it caused kinetic effects comparable to a significant attack through traditional means.

Ch. Ford (then-assistant secretary of state for International security and non-proliferation) (2020) International Security in Cyberspace: New Models for Reducing Risk, *Arms Control and International Security Paper Series*, vol. I, no. 20. https://www.newparadigmsforum.com/p2818

Cyber attacks on critical infrastructure

2021: malware attack on Colonial Oil Pipeline, providing 45% of gas, diesel, and jet fuel for US East Coast

2017: Triton malware in a Saudi petrochemical plant, allowing toxic leaks and explosions

2016: Ukrainian power facility PrykarpattyaOblenergo

Widespread skepticism that cyber capabilities enhance the ability of states to launch highly destructive attacks

Is AI a game changer?

- Security professionals and black hats still work ***artisanally***.
- AI in cyber kill chain may increase speed, destructiveness, autonomy
  - vulnerability identification, tools for attack delivery, exploration of environment, taking control on penetrated systems

Cyber attack surface in the nuclear complex

H. Lin (2021), *Cyber Threats and Nuclear Weapons,* Stanford UP

- *Embedded computing in nuclear delivery systems*
- *Contractor supply chain: off-the-shelf hardware and software available to hackers for examination*
- *Software of Ballistic Missile systems*
- *Nuclear planning and situational awareness*
- *Nuclear decision making and force direction authentication*

Strategic attacks on NC3

Significant non-nuclear strategic attacks include… attacks on the US, allied, or partner civilian population or ***infrastructure***, and attacks on US or allied ***nuclear forces, their command and control, or warning and attack assessment capabilities***.

2018 US Nuclear Posture Review

Summary

1. AI in NC3 and its vulnerabilities
2. AI and cyber attacks to nuclear forces
3. **AI and deterrence**
4. Concluding remarks

AI information warfare synthetic data generation

- Deep fakes erode political leaders’ credibility and fuel misperception of nuclear threats
- Political leaders' credibility, consistency and rationality about second strike response is put into question

AI finders for nuclear hiders

- Orca: autonomous XLUUV carrying out operations for months.
- Identify submarines at chokepoints or emerging from port and pursue them autonomously.
- «Future technologies will make the oceans broadly transparent. Counter-detection technologies will not have the same salience»
  - *Transparent Oceans?* ANU National Security College Publication 2020; *Will the Atlantic become transparent?* British Pugwash report 2016

Summary

1. AI in NC3 and its vulnerabilities
2. AI and cyber attacks to nuclear forces
3. AI and deterrence
4. **Concluding remarks**

A seemingly prudential approach

- Officials from the Office of the Under Secretary of Defense for Research and Engineering told us that DOD is just starting to explore deep learning neural networks, but does not currently have any in use
  - AI Status of Developing and Acquiring Capabilities for Weapon Systems, report GAO-22-104765, 17 february 2022
    - https://www.gao.gov/products/gao-22-104765

AI and arms control

- AI's potential enhancement of verification regimes
- AI integrated in intelligence, surveillance and reconnaissance systems, to strengthen transparency through compliance monitoring of nuclear arsenals and treaty verification.

New commitments and actions?

- The US “should clearly and publicly affirm existing U.S. policy that only human beings can authorize employment of nuclear weapons…
- The US should include such an affirmation in the DoD’s next Nuclear Posture Review and seek similar commitments from Russia and China” (NSCAI 2021, p. 10, 98)
- Prohibit (AI-enhanced) cyber attacks on critical infrastructures, including nuclear weapons and support systems.

Principles, commitments, actions?

- Establish venues to discuss AI's impact on crisis stability and develop trust and confidence building measures
- Get computer scientists involved in discussing impact of AI on nuclear stability
- Overall, emerging AI threats and vulnerabilities add new motives for nuclear non-proliferation and disarmament

Thank you
for your attention!

AWS: AI-powered autonomous weapons systems

- Have the potential to give large conventional military advantages to adopters
- If autonomous systems tilt the conventional military balance, a nuclear-armed adversary may feel incentivized to threat the use of nuclear weapons to avoid military defeat

Beyond AI in NC3
Additional AI threats to nuclear stability

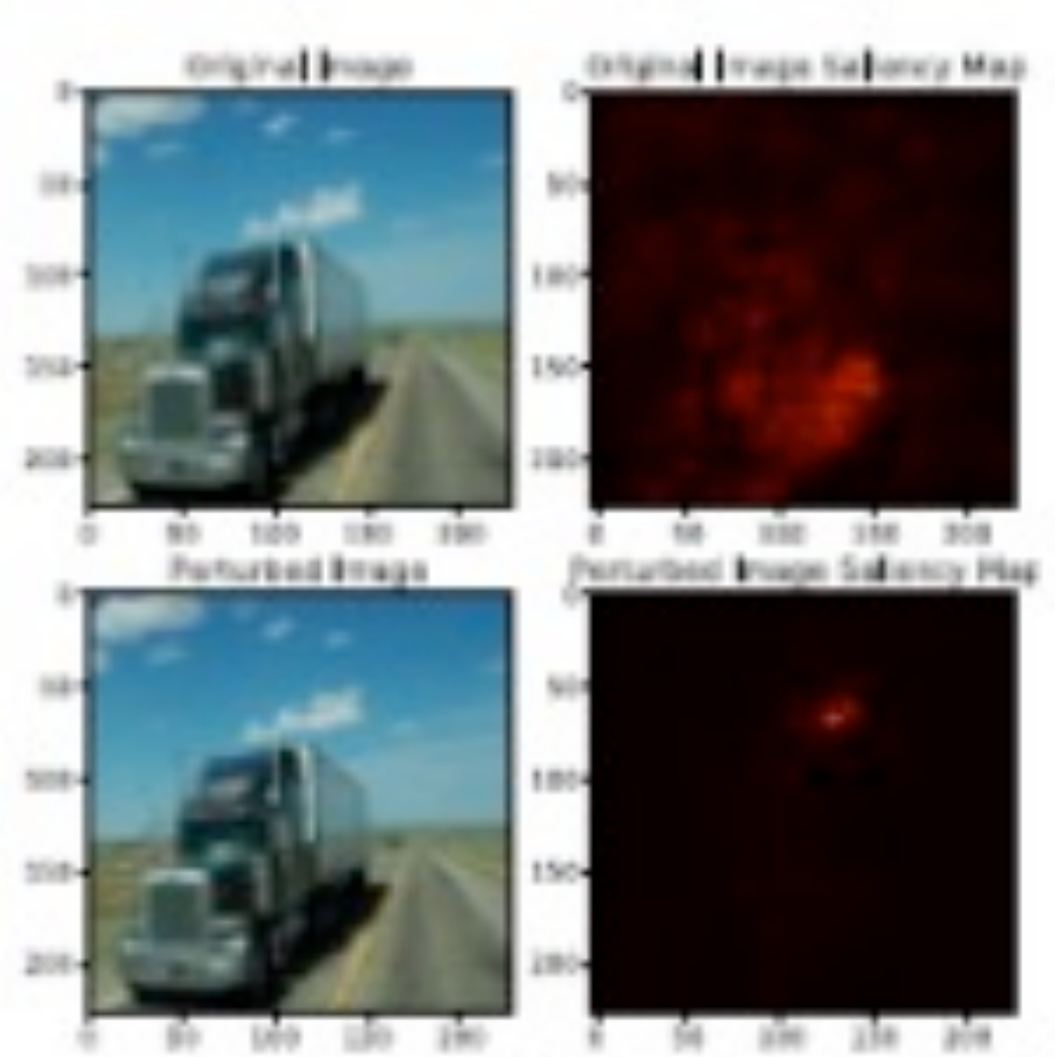
- AI and autonomous (underwater) vessels
  - AUVs trail submarines carrying SLBMs and undermine their stealth.
- AI information warfare
  - Deep fakes erode political leaders' credibility, fueling misperception of nuclear threats and second-strike posture
- AI in cyberwarfare
  - may increase speed and destructiveness of cyberattacks (to NC3)
- AI-powered autonomous weapons
  - have the potential to give large conventional military advantages to adopters and tilt the conventional military balance, incentivizing use of nuclear weapons to avoid military defeat

Induced misclassifications of AI learning systems

- 2017 Misclassification of traffic signs by AI algorithms
- 2018 : Misclassification of medical abnormalities detected by AI systems
- 2018 Misclassification of face images, which can lead to authentication bypass in various scenarios
- 2019: Poisoning data to AI algorithms, resulting in wrong recommendations
- 2020: Adversarial network traffic generation to bypass the security of AI-powered network intrusion detection systems
    - Yamin M. M. et al 2021, Weaponized AI for cyber attacks, J. Information Security and Applications 57, https://doi.org/10.1016/j.jisa.2020.102722
    - Abaimov S, Martellini M. (2022), *Machine Learning for Cyber Agents. Attack and defence*, Springer, Cham.

Adversarial explanations

- The image of a truck on the road was slightly manipulated without affecting the correct classification of the attacked AI system. The changed input induced the system to explain its classification solely in terms of the cloudy sky in the background. No salient features of the truck appearing in the foreground were mentioned in this explanation



- A. Ghorbani, A. Abid, and J. Zou (2019), Interpretation of Neural Networks Is Fragile. Proceedings of the AAAI Conference on Artificial Intelligence, 33(01), 3681-3688

AI and catalytic nuclear war

- Technological and economic ease of AI acquisition and use by non-nuclear third parties for malicious purposes.
- reaction to fake nuclear attack induced by third parties leads to escalating tensions or outright conflict between nuclear powers.
- third party hacks autonomous systems fielded by one nuclear power and use them to expand the scope of the conflict, for instance, by targeting strategic assets.

realistic prospects for AI-powered cyberthreats?

- DARPA Cyber Grand Challenge 2016:
  - AI systems competing for automatic vulnerability finding, patching, exploit generation, cyber security strategy development
  - the performance of the winner of the competition — CMU's ForAllSecure 'Mayhem' system — was significantly weaker than the performance of human specialists
    - p. 49 of https://ccdcoe.org/library/publications/autonomous-cyber-capabilities-under-international-law-2/

Concluding remarks on AI

- AI is a pervasive and malleable technology
- Opacity, statistical correlations without logical or causal reasoning
- Data misclassification (adversarial/inadvertent)
- Synthetic data generation
- AI-powered autonomous systems
- Impact on nuclear stability?