Militarizzazione dell’IA e armi autonome

un punto di vista etico

Un punto di vista etico

- Pace
- Immunità dei non belligeranti e norme della guerra giusta
- Attribuzione di responsabilità (di funzione e retrospettive)
- Minacce al permanere delle civiltà umane

IA: malleabilità e uso duale

scoperta di medicinali e armi di distruzione di massa

- The commercial de novo molecule generator MegaSyn2 is guided by ML model predictions of bioactivity
- MegaSyn2 normally penalizes predicted toxicity and rewards predicted target activity.
- *We simply proposed to invert this logic by using the same approach to design molecules de novo, but now guiding the model to **reward** both toxicity and bioactivity instead*.
- We drove the generative model towards compounds such as the nerve agent VX, one of the most toxic chemical warfare agents developed during the twentieth century
- In less than 6 hours after starting on our in-house server, our model generated 40,000 molecules that scored within our desired threshold. In the process, the AI designed not only VX, but also many other known chemical warfare agents
    - Nature Machine Intelligence 2022 https://www.nature.com/articles/s42256-022-00465-9

Tecnologie IA e loro potenzialità

- L'IA basata sull'apprendimento automatico è una tecnologia malleabile
- Con dati in quantità e qualità sufficienti, l'IA può imparare a eseguire i compiti «intelligenti» più svariati
- Nella Difesa
  - cosa: previsione, pianificazione, diagnosi, generazione di soluzioni, riconoscimento e classificazione, controllo di sistemi ibridi…
  - dove: procurement e logistica militare, sviluppo e operazione di nuovi sistemi d'arma, sorveglianza e intelligence, comando e controllo…

IA e competizione militare

(una corsa alle armi?)

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)

Un punto di vista etico su :

- IA e supporto alle funzioni di comando e controllo (C2)
- IA e autonomia dei sistemi d’arma
- IA e ciberconflitti
- Osservazioni conclusive

IA e supporto alle funzioni di comando e controllo (C2)

IA e C2

- Il sistema IA di «supporto alle decisioni» Habsora (il Vangelo) e Lavender dell'IDF generano liste di obiettivi (edifici o persone) da colpire nella striscia di Gaza
- "during the first 35 days of the war Israel attacked 15,000 targets in Gaza …from 50 targets a year to 100 targets a day"
  - https://www.theguardian.com/world/2023/dec/01/the-gospel-how-israel-uses-ai-to-select-bombing-targets
- "We prepare the targets automatically and work according to a checklist," a source who previously worked in the target division told +972/Local Call. "It really is like a factory. We work quickly and **there is no time** to delve deep into the target. The view is that we are judged according to how many targets we manage to generate." (+972 e Local Call)
  - https://www.972mag.com/mass-assassination-factory-israel-calculated-bombing-gaza/
  - https://www.972mag.com/lavender-ai-israeli-army-gaza/

Controllo umano significativo (CUS) sull'IA?

- CUS implica una configurazione delle interazioni umano-macchina tale che
- (a) la macchina risponde nell'ambiente agli obiettivi e alle ragioni di coloro coinvolti nella sua progettazione e rilascio, pianificazione di azioni e supervisione
- (b) la macchina consente di ricondurre il risultato delle sue azioni alle intenzioni di almeno un agente umano coinvolto nella sua progettazione e rilascio, pianificazione e supervisione delle azioni

Ostacoli al CUS?

- **Tempo:** la macchina produce un numero esorbitante di obiettivi, a fronte di una capacità umana limitata di elaborare e verificare i suggerimenti forniti dalla macchina
- **Pressione psicologica:** risulta dallo sprone ad aumentare la produttività esercitato da superiori, politici e società, a detrimento della qualità del controllo umano
- **Opacità:** dei processi di elaborazione, che ostacola il dialogo con la macchina in base a domande-perché: “perché suggerisci X invece di Y?”
- **Intelligenza aliena:** L’IA individua correlazioni tra i dati di input e di output che sfuggono agli esseri umani e alla loro capacità cognitiva. Questo punto di forza può diventare una debolezza per il controllo, aggravando il problema dell’opacità e della spiegazione.

Ostacoli al CUS?

- **Errore statistico** intrinseco alle decisioni dell'IA. C'è una percentuale di errore accettabile su un falso positivo di attacco nemico o di un obiettivo da attaccare? Il 10%? E' ammissibile sapere in anticipo che ci sarà il 10% di errori? Che possono verificarsi errori più rari ma «gravi»?
- **regolarità del mondo** : l'IA proietta sul futuro regolarità trovate nei dati raccolti in passato. Questa aspettativa funziona/non funziona dove «la sorpresa è essenziale per avere il sopravvento nel momento decisivo» (Clausewitz).

- Quali risposte a questi problemi? XAI, analisi di rischio e risposte prudenziali?

IA e bias da automazione

- Finestre temporali limitate per CUS facilitano il ricorso alle euristiche cognitive (scorciatoie cognitive ***rapide***)
- e l'insorgere di bias da euristica della disponibilità (WISIATI Kahnemann e Tversky), da gregge, da autorità («Chi sono io per contrastare il suggerimento di questa IA?»)

Integrare IA nel comando e controllo nucleare (NC2)?

- **Raccomandazione**: “AI should assist in some aspects of nuclear command and control: *early warning*, early launch detection, and multi-sensor fusion…”
  - (National Security Commission on AI, Final report 2021, p. 104)
- **Motivazione:** “AI may increase reliability, reduce accident risks, shorten processing time, buy more time for decision-makers” (situational awareness) finestra temporale per decidere: 0-20 min
- **Obiezione:**
  - richiesta di “situational awareness” dell’operatore umano
  - fragilità, vulnerabilità, opacità dell’IA

Automazione e falsi positivi: USA 9/11/79

- imminent nuclear attack detected through the early-warning system of NORAD (North American Aerospace Defense Command). National Security Advisor Zbigniew Brzezinski, just one minute before he planned to call President Jimmy Carter to recommend an immediate U.S. nuclear retaliatory response, was informed that the NORAD message was a false alarm caused by software simulating a Soviet missile attack that was inexplicably transferred into the live warning system at the command's headquarters.
  - https://www.armscontrol.org/act/2019-12/focus/nuclear-false-warnings-risk-catastrophe

Automazione e falsi positivi : URSS 26/9/83

- Il sistema sovietico di allerta precoce OKO scambiò dei riflessi di luce solare sulle nuvole per segni distintivi di 5 missili balistici intercontinentali.
- Il colonnello Stanislav Evgrafovič Petrov :“se si inizia una guerra, non lo si fa solo con 5 missili.”
- Buon senso, spiegazioni, ragionamento causale in assenza di dati per addestramento
  - https://www.armscontrol.org/act/2019-12/focus/nuclear-false-warnings-risk-catastrophe

IA e autonomia dei sistemi d’arma

Confronto aereo a distanza ravvicinata tra caccia 15 aprile 2024

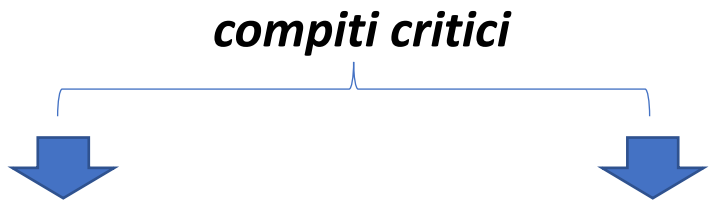
- **US Air Force stages dogfights with AI-flown fighter jet**
- **Two pilots were in VISTA's cockpit to monitor its systems,…but they never had to take over flying.**
- **the lessons learned could apply to more than just dogfighting… to create UAVs that can autonomously fly alongside crewed fighters, carrying out missions such as airstrikes and reconnaissance operations**
    - https://www.defensenews.com/air/2024/04/19/us-air-force-stages-dogfights-with-ai-flown-fighter-jet/

armi autonome AWS

Alle origini del dibattito etico-politico sulla militarizzazione dell'IA

Proprietà salienti delle armi autonome

Un sistema d'arma è autonomo solo se



"una volta attivato, è capace di **selezionare** and **attaccare un obiettivo** senza ulteriori interventi da parte degli esseri umani."

DoD DIRECTIVE 3000.09/2012: AUTONOMY IN WEAPONS SYSTEMS, pp. 13–14.
www.dtic.mil/whs/directives/corres/pdf/300009p

il dibattito etico deontologico e consequenzialista sulle armi autonome (AWS)

- violazioni del DIU (Diritto internazionale umanitario)
- vuoti di responsabilità per atti assimilabili a crimini di guerra
- violazione della dignità umana
- soglia più bassa per avviare nuovi conflitti
- Accelerazione ritmo dei conflitti oltre capacità umane

- rischio più contenuto per il personale militare
- precisione maggiore e tattiche belliche più conservative
- riduzione potenziale del numero di vittime

AA ed etica dei doveri

- Diritto umanitario in guerra (distinzione e proporzionalità)
  - Le AA sanno distinguere persone e cose protette? Sanno valutare la proporzionalità di un attacco?
- Responsabilità
  - Chi è responsabile di ciò che fa una AA? Le difficoltà di previsione sono scusanti per atti materialmente equivalenti a crimini di guerra?
- Dignità umana
  - La dignità umana è negata se “la vittima di una AA non può fare appello all’umanità di ***qualcuno*** che si trovi dall’altra parte”?

AA ed etica delle conseguenze attese

- Pro AA: Conseguenze locali
  - le AA saranno più conservative nelle decisioni di attacco, precise ed esenti da fattori umani invalidanti sui singoli campi di battaglia.
- Contra AA: Conseguenze su larga scala
  - ritmo accelerato dei conflitti, imprevedibilità e incontrollabilità delle interazioni.
  - nuova competizione militare e rischi di destabilizzazione regionale e globale

Piattaforma *differenziata* di ICRC (12.5.2021) per una regolamentazione delle AA

- Un nuovo accordo internazionale vincolante per

1. Proibire le armi autonome *letali (*
2. Proibire le armi autonome *imprevedibili*
3. Regolamentare opportunamente altri tipi di armi autonome
   - limits on the types of target, duration, geographical scope and scale of use, limits on situations of use, requirements for human–machine interaction

- Soluzione prudenziale per il punto 3
  - D. Amoroso & G. Tamburrini (2021), Toward a Normative Model of Meaningful Human Control over Weapons Systems, *Ethics and International Affairs*, 35(2). https://doi.org/10.1017/S0892679421000241

Controllo umano significativo e proposta CICR

- C’è consenso internazionale al CCW sul **controllo umano significativo** (CUS) di *ogni* sistema d’arma
- Il CUS **esclude sempre** l’autonomia nelle funzioni critiche delle AA? Non sempre per il CICR.
- Non c’è una sola taglia di CUS che vada bene per tutte le armi autonome
- Dibattito politico e diplomatico in sede ONU: CCW di Ginevra e Assemblea generale UN.

IA e dissuasione

Dissuasione nucleare

Con il TNP *Trattato di non proliferazione delle armi nucleari* (1970) le potenze nucleari si impegnarono a prevenire la proliferazione delle armi nucleari e ad arrivare gradualmente al disarmo nucleare.

Niente di tutto ciò si è avverato.

Per prevenire un conflitto nucleare, le grandi potenze si affidano ancora alla strategia della dissuasione o deterrenza nucleare.

La triade della dissuasione: missili lanciati da terra, bombardieri dotati di armamenti nucleari, sottomarini dotati di missili balistici a testata nucleare.

Veicoli autonomi e dissuasione

- 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

Deep fakes e dissuasione

- I sindaci di Berlino, Madrid e Vienna hanno fatto video-chiamate nel giugno 2022 con un deepfake del sindaco di Kiev Vitali Klitschko senza rendersi conto dell'inganno
  - https://www.theguardian.com/world/2022/jun/25/european-leaders-deepfake-video-calls-mayor-of-kyiv-vitali-klitschko
- I deep fake erodono la credibilità, la razionalità e la coerenza dei leader politici in relazione alle minacce di risposta a un primo attacco

IA e ciberconflitti

Cyberattacks and the US Nuclear Posture

- 2018 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*. 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.

Cyberattacks and the US Nuclear Posture

- Ch. Ford October 2020 (then-assistant secretary of state for international security and non-proliferation):
- 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. (H. Lin, p. 28)

Ciberattacchi a infrastrutture critiche

- 2017: Triton malware was found in a Saudi petrochemical plant, allowing hackers to take over the plant's security system, possibly leading to explosions and release of toxic gases.
- 2021: malware attack on Colonial Oil Pipeline, providing than 45% of the East Coast's gas, diesel, and jet fuel

- Widespread skepticism that cyber capabilities enhance the ability of states to launch highly destructive attacks.
- Does AI change this appraisal?

AI enhanced cyberthreats?

- Security professionals, white and black hats work artisanally to find and fix/exploit vulnerabilities. The process of finding, countering or exploiting cyber vulnerabilites is mostly ***artisanal***.
- AI for cyberoffence: recognizing target vulnerabilities, creating tools for delivering attacks, installing persistent backdoors, exploring the environment using security breaches, exercising command and control on penetrated systems
- AI may make cyberattacks faster, better targeted, more destructive, more pervasive, more diverse.

Osservazioni conclusive

Ricercatori, controllo degli armamenti, pace

- Impegno di molti fisici per denunciare i pericoli posti dalle armi nucleari, per la riduzione degli arsenali e per il disarmo nucleare
- Manifesto Russell- Einstein del 1955. Nascita del Pugwash
- In Italia: conferenze Amaldi all'Accademia dei Lincei, attività di USPID, scuole internazionali ISODARCO
- Impegno dei biologi e dei chimici per la messa al bando delle armi biologiche e chimiche
- E gli informatici? E' arrivato il loro turno
  - Whistle-blowing, (in-)formare politici e diplomatici, incoraggiare azioni per frenare la militarizzazione senza vincoli dell'IA,...

IA per la pace

- Il nuovo complesso militare-industriale, le piattaforme e gli utenti
- Sistemi di IA per la verifica dei trattati sulle WMD
- La lezione del progetto Maven e Google

Qualche riferimento bibliografico

- Tamburrini, G. (2024). Artificial Intelligence and large-scale threats to humanity, Proceedings of the Digital Humanism Summer School, Vienna, September 2022
- Tamburrini, G. (2023). Cyber threats, nuclear weapons, and the militarization of AI. *XXII Edoardo Amaldi Conference*, Accademia dei Lincei, Rome, April 6-8, 2022
- Tamburrini, G. (2022). The AI Carbon Footprint and Responsibilities of AI Scientists. *Philosophies,* 7(4). (Open access)
  - https://doi.org/10.3390/philosophies7010004
- D. Amoroso, D. Garcia, G. Tamburrini (2022). The weapon that mistook a school bus for an ostrich, *Science & Diplomacy*, AAAS Center for Science Diplomacy. (Open access)
  - https://www.sciencediplomacy.org/article/2022/weapon-mistook-school-bus-for-ostrich
- D. Amoroso and G. Tamburrini (2021). Toward a normative model of meaningful human control over weapon systems, *Ethics & International Affairs* 35(2), (Open access)
  - https://www.cambridge.org/core/journals/ethics-and-international-affairs/article/toward-a-normative-model-of-meaningful-human-control-over-weapons-systems/A3FD9EC4CBD6EA77439211537B94A444

Appunti

- Rand AI and multi domain integration military

Opacità dei processi e il “dar conto”: ragionamento, scienza, etica

- Norme logiche
  - Produzione e validazione di argomentazioni corrette
- Norme scientifiche
  - argomentazioni logiche, causali e statistiche conformi al metodo scientifico
- Norme morali
  - Spiegazione e giustificazione di decisioni morali

- l programma di ricerca della XAI (eXplainable AI)