Panel discussion: Role of AI in conflicts and nuclear risks

The dark side of CyberWorld

Gian Piero Siroli, Physics & Astronomy Dept. Univ. of Bologna & CERN

USPID – October 2023

ICT is intrinsically a dual use technology

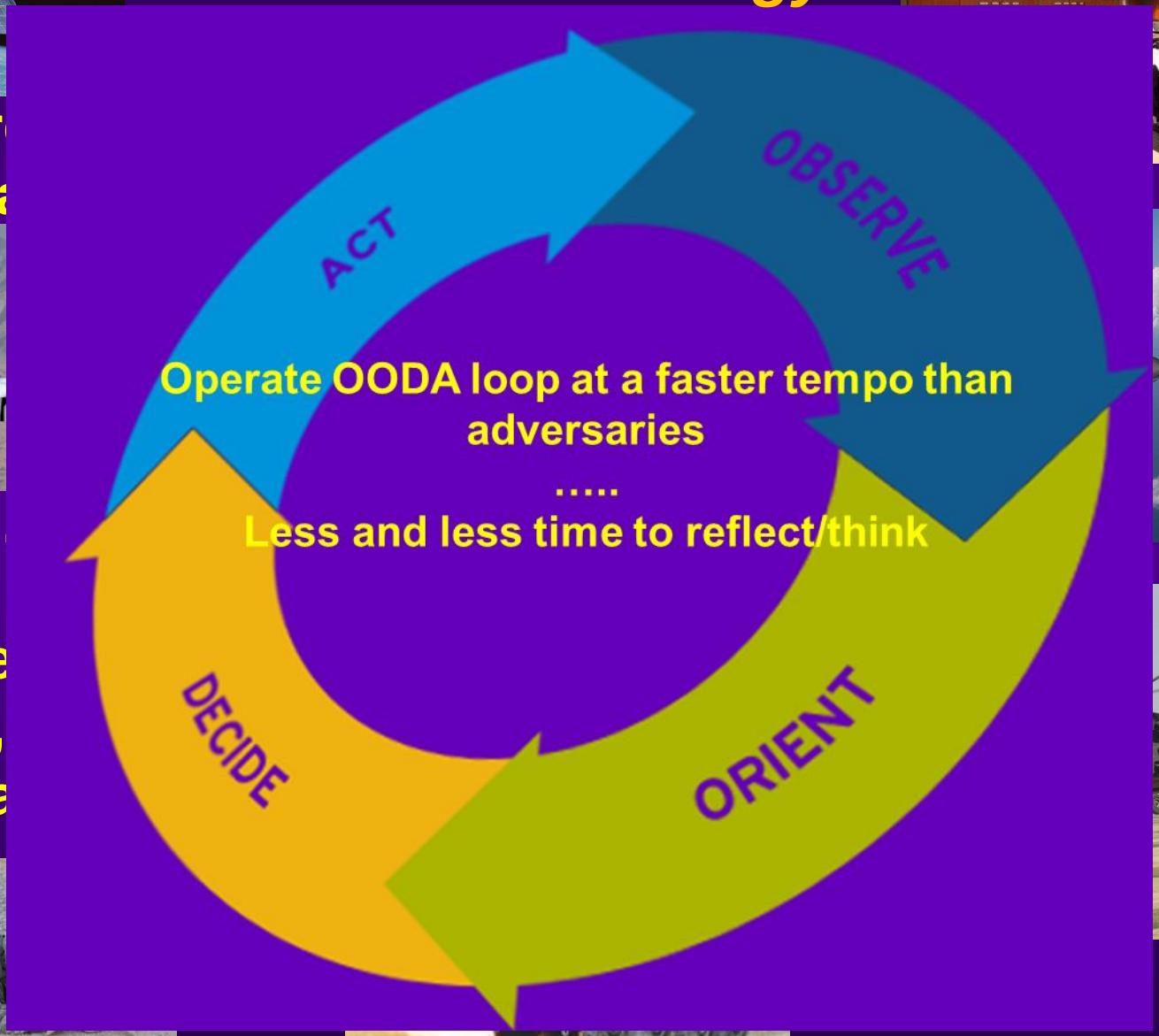
Increasing integration of warfare domains via battlefield digitization

Command, Control, Communications, mission planning & management, interconnection of sensors (Big Data, clouds) across air, land, sea, space domains/assets. Future "Battlefield IoT" environment?!

ICT is intrinsically a dual use technology

Incr
doma

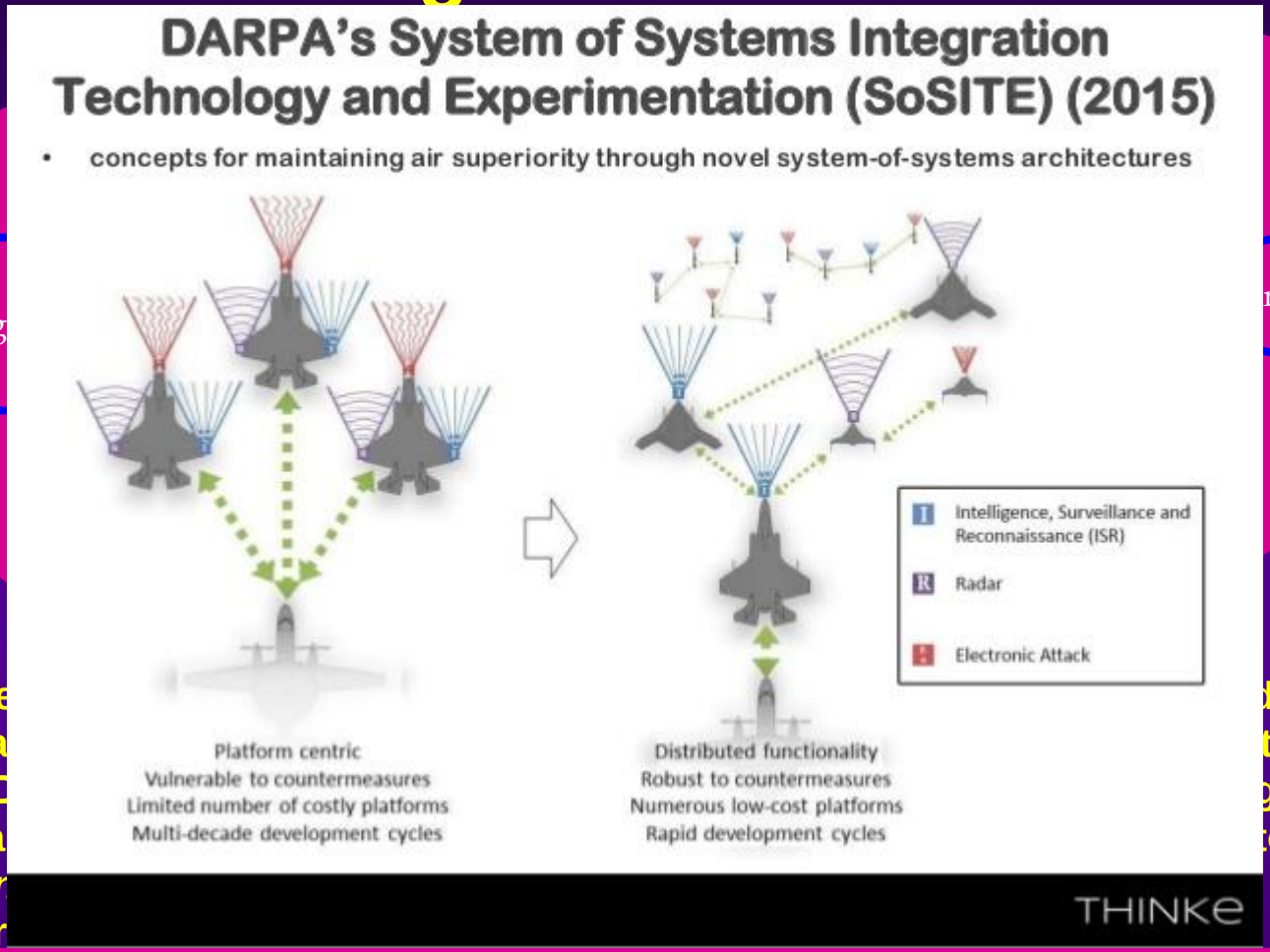
Command
mission
interconne
across air,
Future "Ba



Artificial Intelligence – Machine Learning

- AI - defense of critical Networks: real time, pattern recognition, anomaly detection
- ML - algorithms to efficiently respond to potential network threats in real time
- "Big Data": acquisition, storage, analysis, transfer, visualization, querying, privacy
- Human *in*, *on, out* of the loop? (remote ctrl, semi-autonomous, autonomous). "*Meaningful*" human control
- Cyber-intelligence? Current algorithms not capable of human level reasoning. Presently employed to process & manage (sensor) data, monitor systems integrity, support vocal commands, navigate
- Support C4ISTAR system: Command, Control, Communications, Computing, Information, Intelligence, Surveillance, Targeting Acquisition & Reconnaissance
- Use of AI in the military domain also includes human decision-making support & autonomous systems (LAWS). Probably at an advanced exploratory(?) phase

Artificial Intelligence – Machine Learning



- AI - de… …detection
- ML - a… …time
- "Big D… …g, privacy
- Huma… …onomous).
  "*Mear…*
- Cyber… …reasoning.

We already have weapons that can use AI to search, select and engage targets in specific situations

- Use of AI in the military domain also includes human decision-making support & autonomous systems (LAWS). Probably at an advanced exploratory(?) phase

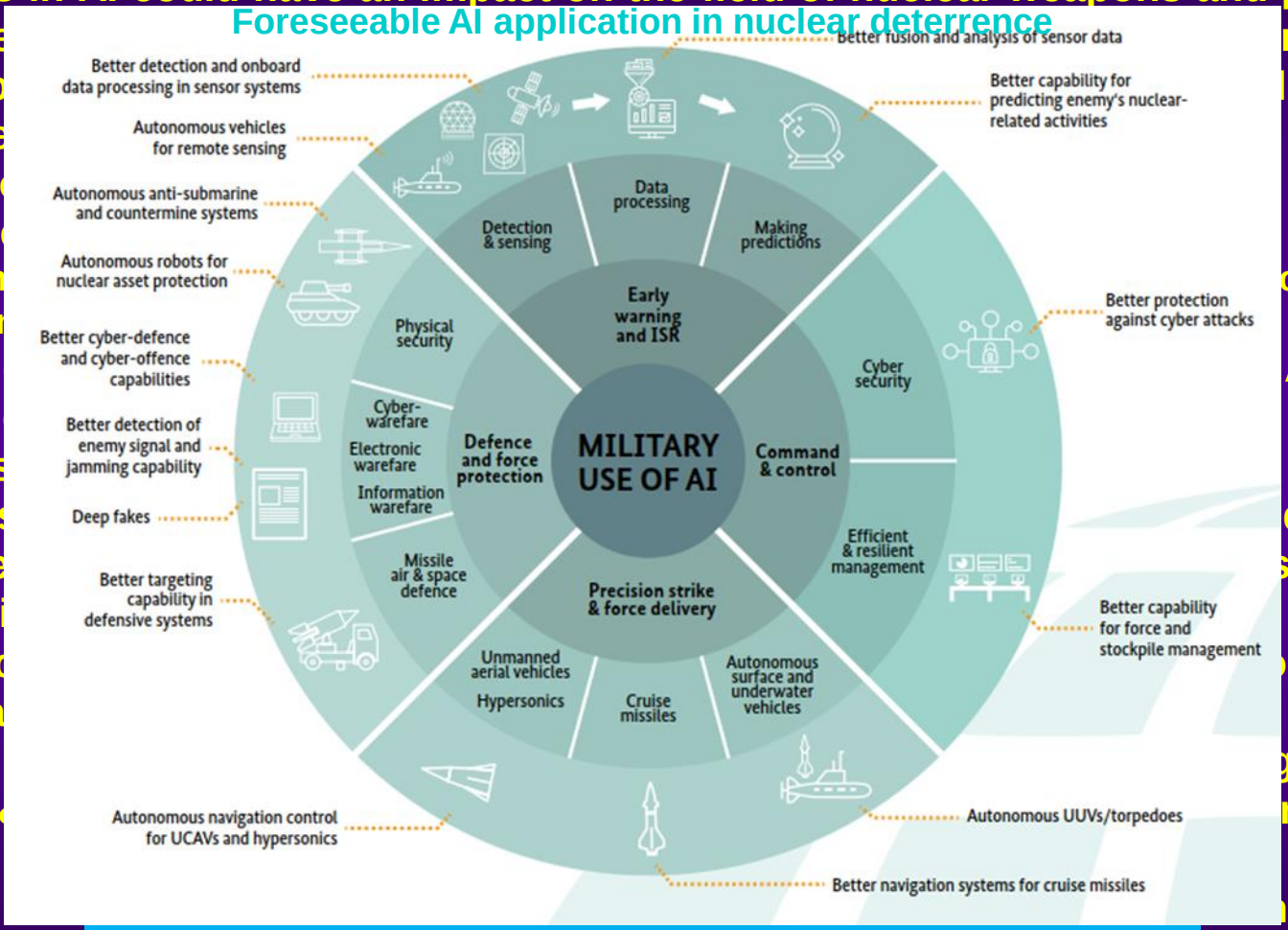
AI and nuclear risk

- Advances in AI could have an impact on the field of nuclear weapons and posture, with consequences for strategic stability and nuclear risk reduction. Nuclear-armed states & international organizations must consider a spectrum of options to deal with the challenges generated by AI
- ML & autonomy could unlock new possibilities for a wide array of nuclear force-related capabilities (early warning, C&C, weapon delivery)
  - Machine Learning (ML) enables development of increasingly capable AI applications
  - Autonomy refers to ability of a machine to execute tasks without human input
- Clear evidence that all nuclear-armed states have made the pursuit of AI a priority. Stabilizing/destabilizing effects?
- Examples:
  - Investment in AI (even non-nuclear-related) by the adversary could threaten a state's future second-strike capability, generating insecurity, decreasing strategic stability and increasing risk of a nuclear conflict
  - AI could fail or be misused in ways that could trigger an accidental or inadvertent escalation of a crisis or conflict into a nuclear conflict
- Support awareness-raising measures helping relevant stakeholders (governments, industry & civil society) to understand a realistic challenges posed by AI in the nuclear arena
- Support transparency & CBMs to reduce misperception/misunderstanding among nuclear-armed state
- Discuss and agree on concrete limits to the use of AI in nuclear forces/infrastructures

AI and nuclear risk

- Advances in AI could have an impact on the field of nuclear weapons and posture, with conseque… med states & internatio… leal with the challenge…
- ML & aut… force-related capabiliti…
  - Mach… ons
  - Auto…
- Clear evi… AI a priority. Stabilizin…
- Examples…
  - Inves… d threaten a state… sing strategic stabi…
  - AI co… r inadvertent esca…
- Support … governments, industry … n the nuclear arena
- Support … ding among nuclear-arme…
- Discuss and agree on concrete limits to the use of AI in nuclear forces/infrastructures

Foreseeable AI application in nuclear deterrence



V.Boulanin et al "Artificial Intelligence, Strategic Stability and Nuclear Risk" (SIPRI 2020)

AI “weaponization”

***Weaponized AI*: malicious AI mechanisms can degrade the performance and disrupt the normal functions of benign AI algorithms, while providing technological edge attack scenarios in both cyberspace and physical spaces**

- **Deeplocker (IBM Research 2018): AI-based malware triggering mechanisms. DeepLocker's Deep Neural Network model provides “trigger conditions” to be met for malware activation. In case the target is not found, malware stays obfuscated inside the app, making reverse-engineering an almost impossible task - DeepLocker - Concealing Targeted Attacks with AI Locksmithing - Black Hat USA 2018 (slides). DeepLocker: How AI Can Power a Stealthy New Breed of Malware**

- **EvilModel: Hiding malware inside Neural Networks (proof-of-concept 2021)**
  Deliver malware covertly and evasively through a Neural Network by embedding malware in neurons with minor or no impact on the performance of NN and overcoming AV engines. With widespread application of AI, utilizing NN may become a forwarding trend of malware - Neural networks can hide malware, researchers find - TechTalks

- **Introductory non-technical primer: The Weaponization of Increasingly Autonomous Technologies: Artificial Intelligence (UNIDIR 2018)**
- **Human-Machine Interfaces in Autonomous Weapon Systems (UNIDIR 2022)**
- **Confidence-Building Measures for Artificial Intelligence: A Framing Paper (UNIDIR 2022)**

More on AI: LLM (Large Language Model)

What about asking ChatGPT how to build a bomb? Ethical filters (AI ethics) - Alignment of LM: process of ensuring that models produce outputs consistent with human intentions & values
Challenges/key aspects:

- **Data quality & diversity: training step on large-scale datasets containing noisy, biased, outdated info, affecting accuracy, reliability, fairness of outputs ("garbage in – garbage out")**
- **Human supervision & feedback: human feedback to fine-tune/steer outputs can be costly, time-consuming, subjective, inconsistent (Principle-Driven Self-Alignment of Language Models from Scratch with Minimal Human Supervision (arxiv.org))**
- **Model transparency & interpretability: LLM often complex & opaque, difficult to understand how outputs/decisions are generated, hindering trust & accountability ("black box")**
- **Adversarial attacks & robustness: Language models may be vulnerable to malicious inputs or manipulations causing undesirable or harmful outputs. This can pose a threat to the safety and security of the models and their users**
- **AI & LLMs sometimes *"hallucinate"* (i.e. outputs may sound plausible but either factually incorrect or unrelated to the given context)**
- **…and more…**

…So, we need ethical filters on LLMs…
BUT

1. **Universal attacks on LLMs ("Universal Jailbreaks"), methods allowing users to bypass LLM ethical filters restrictions by prompt injection. "Universal and Transferable Adversarial Attacks on Aligned Language Models" (Ask Guglielmo)**
2. **Models without ethical filters ("WormGPT: What to know about ChatGPT's malicious cousin" ZDNET). Open-source models**
3. **You can always (re)train your own model, if have enough CPU/GPU power**

More on AI: LLM (Large Language Model)

What about asking ChatGPT how to build a bomb? Ethical filters (AI ethics) - Alignment of LM: process of ensuring that models produce outputs consistent with human intentions & values
Challenges/key aspects:

- **Data quality & diversity: training step on large-scale datasets containing noisy, biased,**

Is all this safe?
Can we control this technology?
(meaningful human control)
Do we trust AI? Or maybe the real problem…
…is that…AI trusts us?!
Can you see a way-out?
Did we inadvertently open another Pandora box??

…So, we need ethical filters on LLMs…
BUT

1. **Universal attacks on LLMs ("Universal Jailbreaks"), methods allowing users to bypass LLM ethical filters restrictions by prompt injection. "Universal and Transferable Adversarial Attacks on Aligned Language Models" (Ask Guglielmo)**
2. **Models without ethical filters ("WormGPT: What to know about ChatGPT's malicious cousin" ZDNET). Open-source models**
3. **You can always (re)train your own model, if have enough CPU/GPU power**

Food for thought

- The AI Power Paradox: Can States Learn to Govern Artificial Intelligence - Before It's Too Late? (Foreign Affairs, 2023)
- Why artificial intelligence is now a primary concern for Henry Kissinger (The Washington Post, 2022)
- The World Economic Forum wants to develop global rules for AI (MIT Technology Review, 2019)
- Musk, Hawking warn of 'inevitable' killer robot arms race (Wired UK, 2015)
- ChatGPT (MS Bing Chat), Bard (Google), LLaMA-2 (Meta), Claude etc. The main ones are currently developed by very large ICT corporations, owning large computing infrastructures/clouds

- ChatGPT rival WormGPT with 'no ethical boundaries' sold to hackers on dark web - Europol warns AI tool is 'extremely useful' for cyber criminals (The Independent, 2023)

H100, Nvidia's latest GPU optimized to handle large artificial intelligence models, which is used to create text, computer code, images, video or audio. (Nvidia/Handout/Reuters)

Cyber diplomacy

Going forward:

in 2018 established two parallel (hopefully converging & complementary) processes to discuss ICT security in 2019-2021 - Outcome: VERY difficult & hard convergence reached in 2021

- **Open-Ended Working Group (OEWG) open to all Member States (chair Amb.Jürg Lauber CH). Report to GA in 2020. OEWG holding an inter-sessional consultative meeting with industry, civil society, NGOs and academia. New wider *multistakeholder approach* ("The Value of Multistakeholder Engagement")**
- **GGE of 25 members (chaired Amb.Guilherme de Aguiar Patriota BR). Final report in 2021. Chair holding consultations with the wider membership in between sessions. Consultations with regional organizations (AU, EU, OAS, OSCE, ASEAN)**

OEWG should refer to shared conclusions of previous GGEs (2015 A70/174). OEWG represents by itself a sort of CBM

➔ Final report (march 2021)

Current process: OEWG 2021-2025

- **"The right to privacy in the digital age", A/RES/68/167 (2013)**
- **Proposed universal code of conduct for Information Security (2015)**
- **G7 Declaration on responsible state behaviour in cyberspace (2017)**

REAIM 2023 & UN Security Council

(something is finally happening, concerning the military domain)

- **REAIM (Feb 2023): first global Summit on responsible AI in the military domain (The Hague). Platform for all stakeholders to discuss key opportunities, challenges & risks associated with military applications of AI**

- **UN Security Council to hold first talks on AI risks (Reuters, July 2023)**
  First formal discussion on "Artificial intelligence: opportunities and risks for international peace and security".
  In June UN Secretary General backed a proposal by some AI executives for the creation of an international AI watchdog body like IAEA.
- **UN council to hold first meeting on potential threats of AI to global peace (AP)**
  "These scientists and experts have called on the world to act, declaring AI an existential threat to humanity on a par with the risk of nuclear war".

- **➔ UN Secretary-General's remarks to the Security Council on Artificial Intelligence (July 18th, 2023): "…AI has tremendous potential but also major risks about possible use for example in autonomous weapons or in control of nuclear weapons…shocked by the newest form of generative AI, a radical advance in its capabilities (probably a step function in AI evolution)…even its own designers have no idea where their stunning technological breakthrough may lead…consider the impact of AI on peace and security, already raising political, legal, ethical & humanitarian concerns…"**

Quoting a recent movie on a different(?!) context:
"You can lift the rock without being ready for the snake that's revealed"

- **Guterres calls for AI 'that bridges divides', rather than pushing us apart (UN News)**
- **Briefing by UNTV**

Pugwash Conferences on Science and World Affairs

(Founding charter Russel-Einstein Manifesto, 1955)

- Workshops on cyber-security and warfare (Geneva Dec 2018, summary & report). II cyber-workshop (Jan 2020), summary. III cyber-workshop (Mar 2022)
  - International database of national points of contact for addressing cyber security threats and actions
  - Alternative centers and capability of communications and analysis
  - Sharing of Indicators of Compromise (IoC)
  - Publication of vulnerabilities and of incident reporting in cyber-attacks.
  - National and international "Bug-bounty" programs
  - Endorse integrity of encryption protocols
  - Prohibit proliferation of cyber-weapons
  - Reiterated commitment of existing IHL obligations on non-attack of nuclear or critical infrastructures and non military targets
- Workshop sobre Seguridad Informática en América Latina (Bariloche Oct 2019)

Accredited to participate to OEWG meetings (ECOSOC status) to contribute to international norms development process

Role of Pugwash on awareness raising & education on ICT technologies

A personal private "vision" of an international cyber-monitoring system (CBM, situational awareness, international early warning & collaboration under attack, security data sharing…) "CSBMs for the Cyber Realm"

Pugwash Conferences on Science and World Affairs

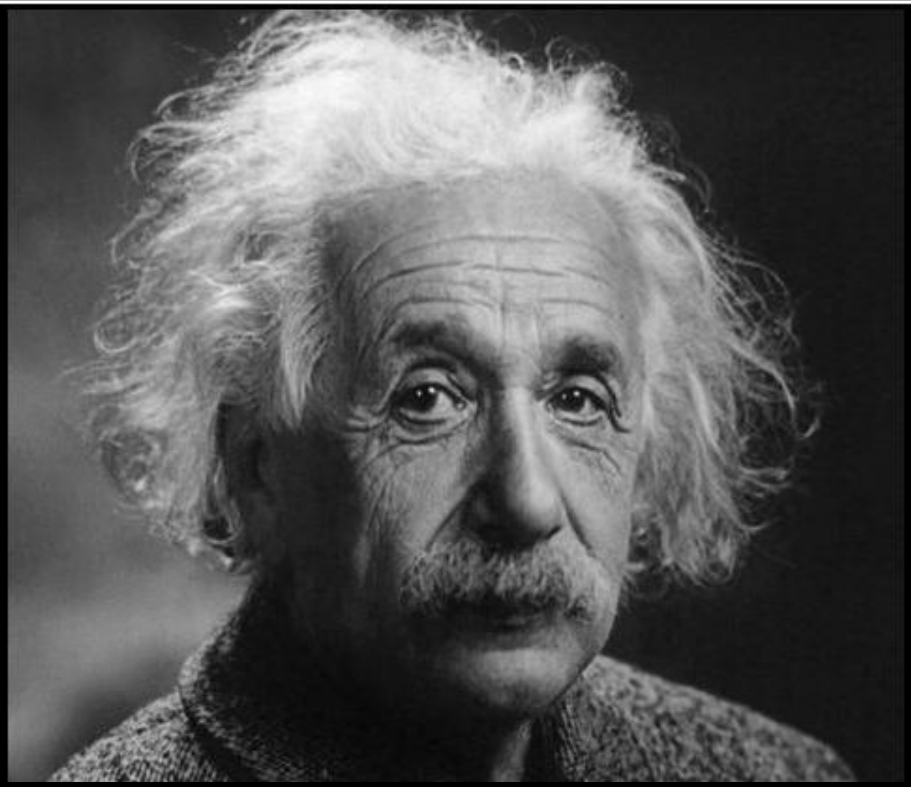
(Founding charter Russel-Einstein Manifesto, 1955)

Role of Pugwash (personal view):

- **Follow UN dynamics in the cyber & info domains through OEWG (collaborate with other stakeholders?) and give input (critical & nuclear infrastructure). Propose CBMs. *Cyber-diplomacy***
- **Raise *awareness* at decision maker & diplomatic level, academia, civil society etc. Promote international cooperation**
- **Organize cyber workshops or any other relevant format, strengthening collaboration in education. *Capacity building***

Probably it's what Pugwash has been doing since foundation but including the "new" cyber/info domain.
A critical mass of people with different backgrounds is needed.

(Personal experience teaching course "Cybersecurity & cyberwar" for master degree in International Relations)



Solution is not at the ICT technical level only

“The importance of securing international peace was recognized by the really great men of former generations. But the technical advances of our times have turned this ethical postulate into a matter of life and death for civilized mankind today, and made it a moral duty to take an active part in the solution of the problem of peace, a duty which no conscientious man can shirk”

Russel-Einstein Manifesto (1955) (Pugwash founding charter)