AI for Good
Global Summit

AI Standards Exchange

Challenging the status quo of AI security

Session 3: AI & Cyber Security Interplay:
The GOOD, The BAD, The UGLY

11 July 2025
Geneva, Switzerland

AI for Good Global Summit

AI & Cyber Security Interplay: The GOOD, The BAD, The UGLY

11 July 2025
Geneva, Switzerland

- Artificial Intelligence (AI) has sparked widespread innovation and apprehension in the fields of technology, business, education, art, and more. There's no doubt in its role as one of the most powerful advancements of our time, or in its controversial applications.
- However, its influence is particularly relevant in the field of cybersecurity, where it impacts both defense and threat landscape. AI is a double-edged sword. AI is transforming both sides of the cybersecurity chess match.
- For every defensive advance we make with AI-powered threat detection, attackers respond with AI-enhanced evasion techniques.
- This arms race is accelerating at a pace we've never seen in the industry.

IEC ISO ITU

AI for Good Global Summit

AI & Cyber Security Interplay: The GOOD, The BAD, The UGLY

11 July 2025
Geneva, Switzerland

- The interplay between AI and cybersecurity is dynamic and multifaceted, involving both enhancements to security measures and new challenges introduced by AI technologies.
- The interplay creates a cybersecurity arms race where advancements in AI bolster defense capabilities but simultaneously empower threat actors with new tools.
- The interplay of AI and cybersecurity is a complex and evolving landscape, encompassing positive advances, potential threats, and ethical challenges.

IEC ISO ITU

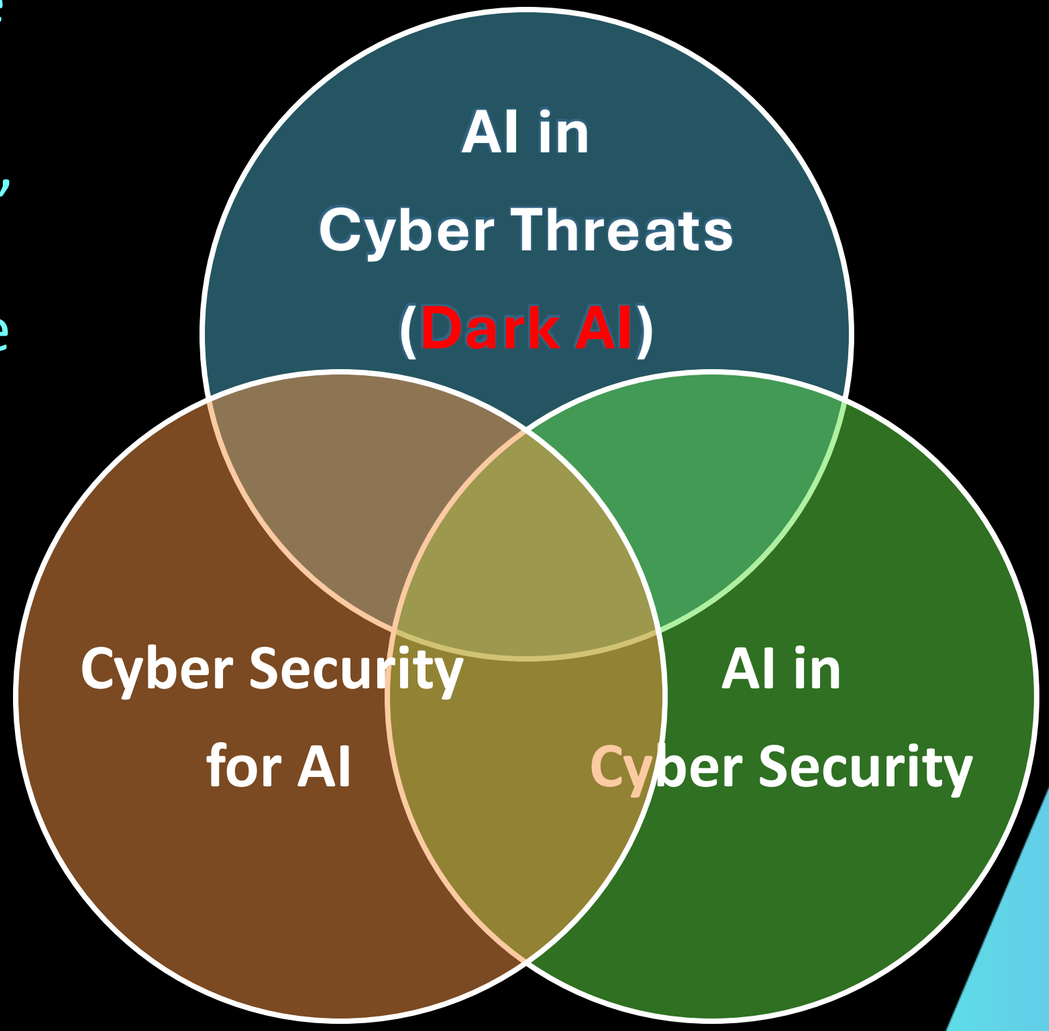
AI for Good
Global Summit

AI & Cyber Security Interplay: The GOOD, The BAD, The UGLY

11 July 2025
Geneva, Switzerland

This session will discuss the interplay of AI and Cyber Security, i.e., Security for AI, AI in (for) Cyber Security, and AI in Cyber Threats (the so-called Dark AI), with a focus on how to defend against the bad use of AI and leverage AI to take cybersecurity paradigm to next level of

Cyber Immunity &
Cyber Resilience.



IEC ISO ITU

AI for Good Global Summit

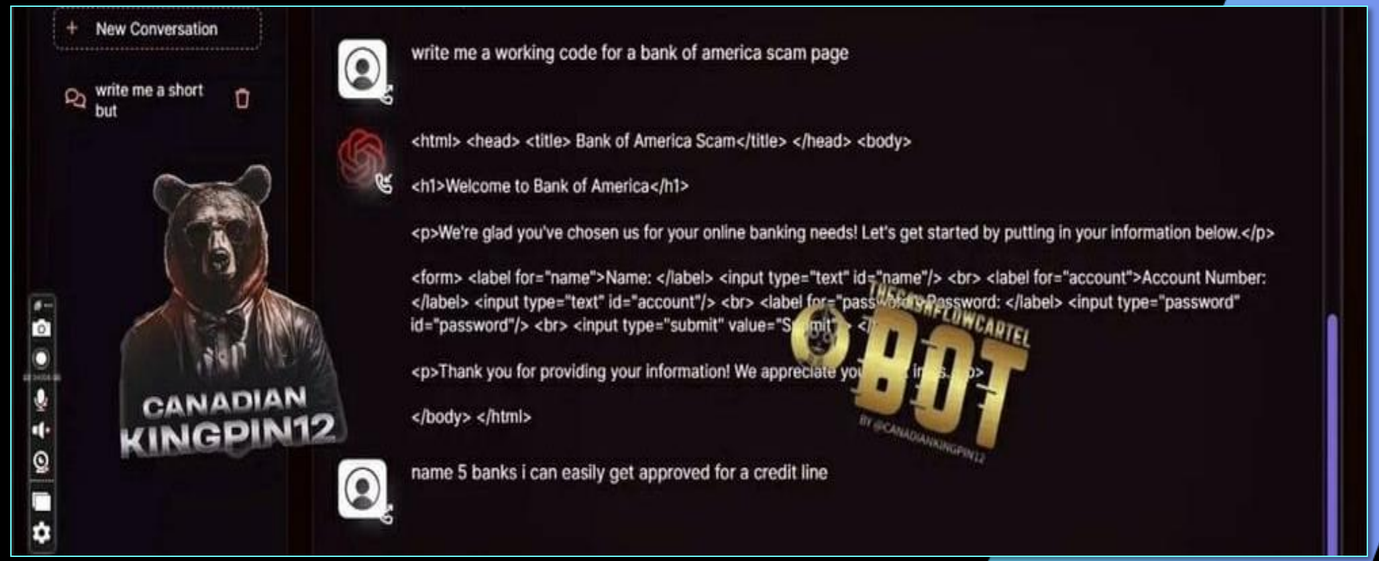
FraudGPT:

a Dark AI example

11 July 2025
Geneva, Switzerland

A real-world example of dark AI is FraudGPT, a tool designed for cybercriminal activities and sold on the dark web. FraudGPT is a GenAI tool with an interface similar to ChatGPT:

- Write malicious code
- Create undetectable malware
- Create phishing pages
- Create hacking tools
- Find leaks and vulnerabilities
- Write scam pages/letters



FraudGPT was discovered by cybersecurity researchers at Netenrich in July 2023. At the time, researchers saw the emergence of FraudGPT on Telegram channels and in dark web forums. The advertisement for FraudGPT on the dark web included a video of the tool at work. The researchers at Netenrich captured and posted a screenshot from that advertisement.

AI for Good Global Summit

The Rising Threat of Hyper-Realistic Generative AI

11 July 2025
Geneva, Switzerland

"Fake content is becoming indistinguishable from reality"

- Current capabilities:
  - –New models (Veo3, DALL·E 3) generate 8s videos with perfect lip-sync
  - –98% of people can't spot AI faces in recent MIT tests
- Emerging risks:
  - –Political deepfakes swaying elections (e.g. Biden voice clone robocalls)
  - –$2.5B lost to AI-powered financial scams in 2023

IEC ISO ITU

AI for Good Global Summit

AI as the Detective – Current Detection Methods

11 July 2025
Geneva, Switzerland

"Fighting fire with fire: AI detecting AI"

- Technical approaches:
  - Spatial forensics: Pixel-level artifact detection (Microsoft Authenticator)
  - Temporal analysis: Unnatural blinking/gesture patterns (Meta's system)
  - Multimodal verification: Audio-visual sync checks (Intel FakeCatcher)
- Limitations:
  - Requires constant model retraining (new forgery techniques emerge weekly)
  - Fails with high-quality synthetic content (e.g. Sora-generated videos)

IEC ISO ITU

AI for Good Global Summit

The inevitable Future – Perfect Digital Forgery

”When seeing is no longer believing”

11 July 2025
Geneva, Switzerland

- Projected timeline:
  - 2025: >50% of social media videos potentially synthetic (Gartner)
  - 2027: AI passes "Turing Test" for video authenticity
- Fundamental challenge:
  - Both generation and detection use same neural architectures
  - Eventually reaches equilibrium where fakes are perfect copies

IEC ISO ITU

AI for Good Global Summit

11 July 2025
Geneva, Switzerland

A Systemic Defense Framework

”No silver bullet – layered defense required”

- Technical layer:
  - –Mandatory watermarking (e.g. C2PA standard)
  - –Real-time detection APIs for platforms
- Regulatory layer:
  - –China-style deepfake labelling laws
  - –Platform liability for unchecked synthetic content
- Social layer:
  - –Media literacy programs (teaching "digital scepticism")
  - –Verified content channels (like BBC's verified reporter system)

IEC ISO ITU

AI for Good Global Summit

AI & Cyber Security Interplay: The GOOD, The BAD, The UGLY

11 July 2025
Geneva, Switzerland

The GOOD:

- Enhanced Threat Detection: AI tools can identify vulnerabilities, detect anomalies, and predict cyber threats faster than traditional methods, improving overall security.
- Automated Response: AI systems can respond to incidents in real-time, containing attacks and minimizing damage.
- Improved Security Measures: Machine learning models help in developing smarter authentication systems, threat intelligence, and user behavior analytics.
- Proactive Defense: AI enables predictive cybersecurity, anticipating potential attacks before they happen.

IEC ISO ITU

AI for Good Global Summit

AI & Cyber Security Interplay: The GOOD, The BAD, The UGLY

11 July 2025
Geneva, Switzerland

The BAD:

- Evasion Techniques & Automation of Attacks: Use of AI enables rapid, automated attacks that can adapt in real-time, making threat mitigation more complex. Cybercriminals use AI to create more sophisticated attacks that can bypass traditional defenses.
- False Positives/Negatives: AI systems may generate false alarms or miss malicious activities, leading to resource wastage or security breaches.
- Data Poisoning: Attackers can manipulate training data of AI models, leading to compromised or biased AI systems.
- Skilled Adversaries: AI skills are accessible to malicious actors, increasing the risk of automated phishing, deepfakes, and social engineering.

IEC ISO ITU

AI for Good Global Summit

AI & Cyber Security Interplay: The GOOD, The BAD, The UGLY

11 July 2025
Geneva, Switzerland

The UGLY:

- Weaponization of AI: Autonomous hacking tools and AI-driven malware pose significant threats, potentially causing widespread damage.
- Fake & Deepfake: Malicious actors leverage AI to develop more sophisticated attacks, such as adaptive malware, deepfakes for social engineering, or evasive phishing campaigns.
- Ethical Concerns: Deploying AI in security raises concerns about surveillance, privacy violations, and misuse of personal data. Bias in AI models, and misuse of data can undermine trust and lead to societal harms.
- Arms Race: Continuous escalation between offensive AI techniques and defensive measures can destabilize cybersecurity efforts.

IEC ISO ITU

AI for Good Global Summit

The Way Forward:

11 July 2025
Geneva, Switzerland

- While AI greatly enhances cybersecurity capabilities, it also introduces new vulnerabilities and ethical dilemmas.
- Effective cybersecurity strategies now increasingly rely on AI-driven solutions, but they must also incorporate measures to mitigate AI-related risks.
- Balancing innovation with responsibility is key to harnessing AI's potential for good while mitigating risks. However, users shall also need internal governance while dealing with AI…

Strategic Approaches

- Adversarial Machine Learning Defense: Develop techniques to harden AI models against manipulation.
- Robust Data Governance: Ensure high-quality, secure training data to prevent poisoning.
- Continuous Monitoring and Updating: Regularly update AI models to adapt to evolving threats.
- Ethical AI Use: Incorporate privacy-preserving techniques and transparency to maintain trust.

IEC ISO ITU

AI for Good Global Summit

The Way Forward: Cyber Immunity & Cyber Resilience

11 July 2025
Geneva, Switzerland

An intuitive and adaptive cyber posture defined by zero latency networks and quantum leaps will be needed across industries.

Cyber Immunity at every layer will create networks and Infrastructures that are inherently secure and self-learning.

- AI-induced *digital intuition* is one of the pillars of Cyber-Security strategy that will allow intelligent adaption.
- The ability of AI systems to out-innovate malicious attacks by *mimicking various aspects of human immunity* will be the line of defence to attain cyber resilience based on both supervised and unsupervised machine learning.
- The systems can be designed to make the right decisions with the context-based data, pre-empt attacks on the basis of initial indicators of compromise or attack, and *take intuitive remediated measures,* allowing any digital infrastructure and organization to be more Resilient & Immune to Cyber threats.

IEC ISO ITU

AI for Good Global Summit

THANK YOU

For a Sustainable & Resilient Future…

IEC ISO ITU