

Weaponizing the AI-Insider Threat Nexus: How Threat Actors Leverage AI

Audrey Crowe

Program Manager
National Security & Critical Infrastructure
Canadian Nuclear Laboratories

June 2026, STUK, Finland



Agenda

1. How AI acts as a Force Multiplier for Threat Actors

- Case Study: DPRK's Illegitimate Remote Workforce

2. How AI acts as Catalyst for Social Engineering

- Case Studies: Arup & Indonesian financial sector

3. How Agentic AI can act as an Insider Threat

- Case Study: AI as a malicious insider
- Case Study: AI as an inadvertent insider

4. Conclusion

5. Discussion

Source: Audrey Crowe (2026). "Weaponizing the AI-Insider Convergence: How Threat Actors Hijack Legitimate Access Pathways". In *Artificial Intelligence and the Security of Civilian Nuclear and Radioactive Materials, Facilities and Activities*. Ed. Sarah Case-Lackner. Springer . [Manuscript Submitted for Publication]



How AI acts as a Force Multiplier for Threat Actors



1. AI – A Force Multiplier for Threat Actors

With broad access to AI tools & extensive open-source information, threat actors benefit from:

Upskilling

- Democratization of advanced knowledge in complex technical fields
 - Coding, programming, nuclear sciences & engineering, industrial processes, operational technology
- Lowers the barrier of entry
 - Technical knowledge available in plain & multiple language

Productivity Tool

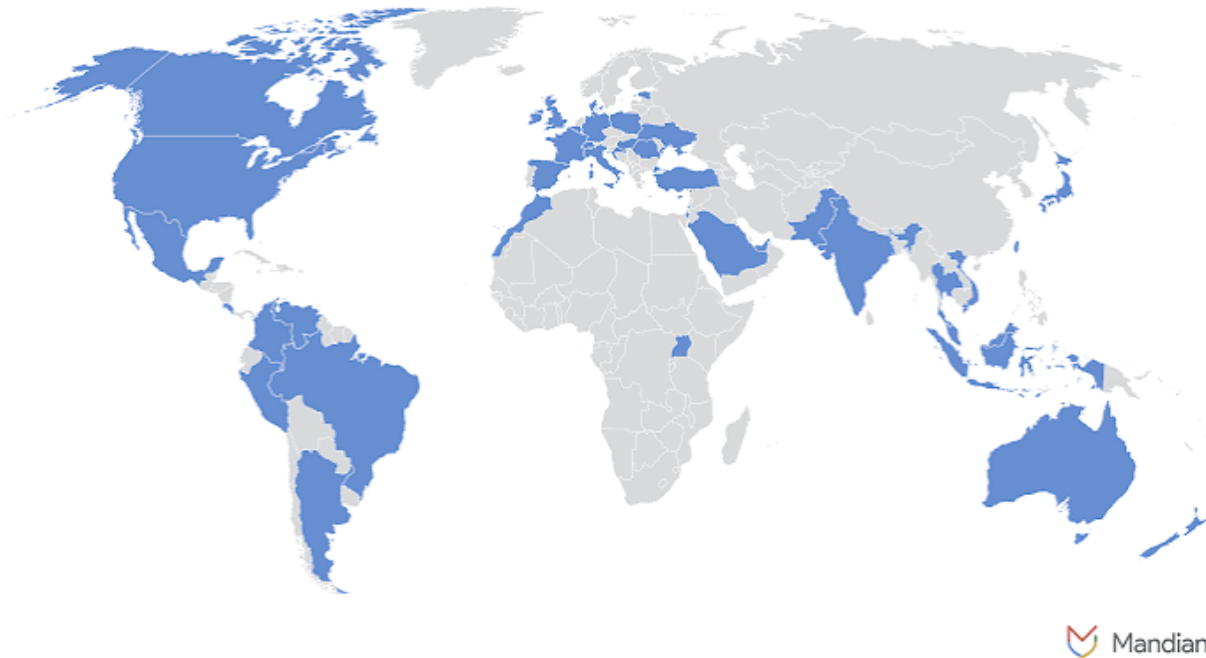
- Automation of repetitive or labour-intensive tasks:
 - Data scraping, aggregation & parsing, vulnerability scanning, credential stuffing
- Analysis & decision-making support:
 - Data classification & prioritization, high-value target selection, attack pathway development

Content Generation & Validation

- Malicious code
- Document, images, audio, video
 - Cloning (deep-fakes)
- Highly-personalized communications
 - Documents, e-mails, verbal or video communication scripts
 - Translation



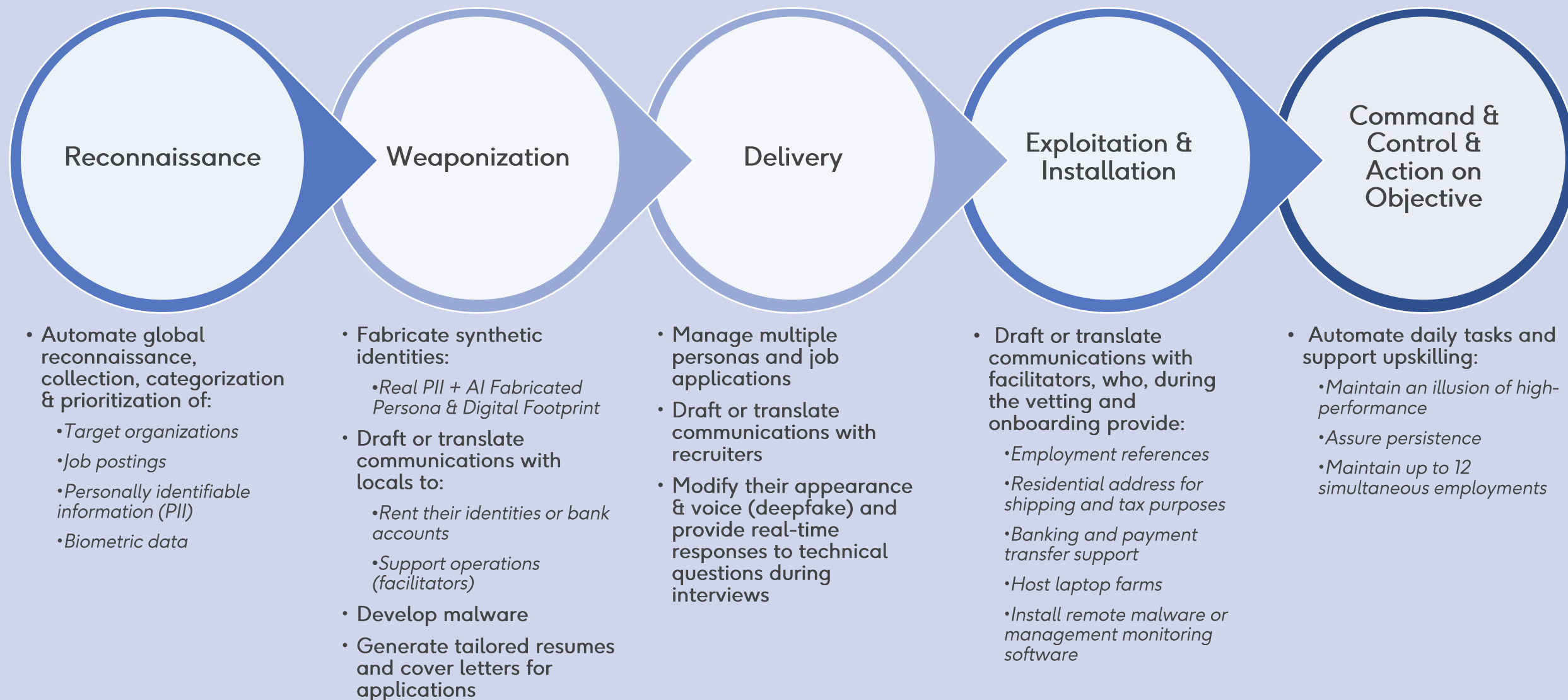
States affected by DPRK's illegitimate remote workforce



When	• Operating since at least 2017 • Upsurge since COVID-19 pandemic
What	• Infiltrated 3,000-10,000 remote operatives as employees – malicious insiders
Where	• Globally, in supply chain: • Professional services & consulting (IT, managed services, cybersecurity), • Software development & emerging technologies companies • Many Fortune 500 companies affected • Found in Google & Amazon talent pools
Why	• Financial Gain • Employment salary, ransomware, post-employment extortion • US\$350-800M/year – 2-5% GDP (£260-595M, €301-689M) • Acquisition of IP & sensitive information • Source code & controlled technologies • Lay of the land, credentials • Conduct of or prepositioning for sabotage operations

Case Study: DPRK's Illegitimate Remote Workforce

How: Extensive use of AI throughout the entire scheme



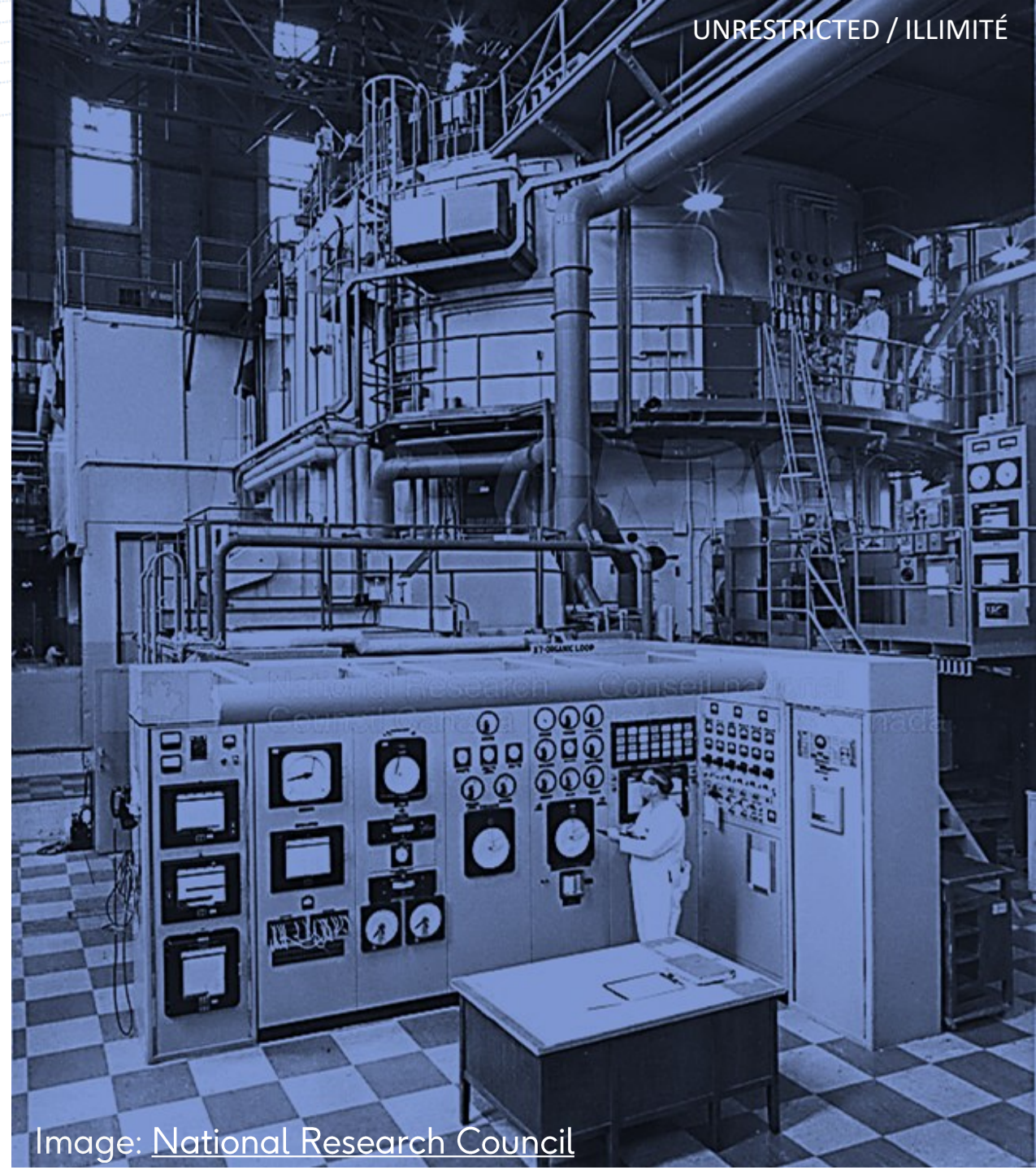
What does this mean for Nuclear OT?

1. Increased malicious activity against nuclear OT

- Adversarial use of AI intensifies the global threat landscape
- Rapid technical upskilling and productivity gains increases the risks of cyberattacks:
 - Threat Actors
 - ↑ number of credible actors
 - ↑ sophistication
 - Threat Activities
 - ↑ frequency
 - ↑ severity
 - ↑ success rate

2. Increased risk of providing threat actors with legitimate access to nuclear OT

- Threat actors make extensive use of AI to infiltrate malicious insiders in:
 - Supply chain
 - Organizations



How AI acts as a Catalyst for Social Engineering

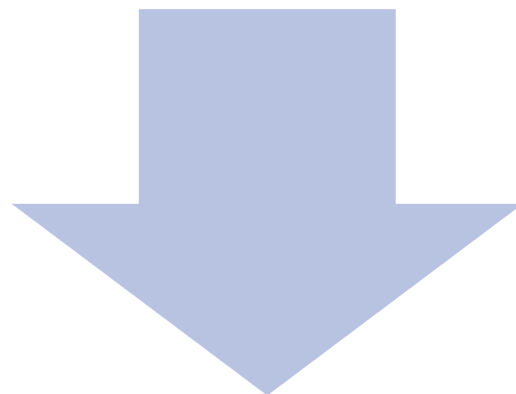


2. AI – A Catalyst for Social Engineering

From 2022 to 2025:



1200% increase in phishing attacks, globally



Time required to draft a realistic phishing e-mail fell from **16 hours** to **5 minutes**

Sources: World Economic Forum, Accenture. (2026). *Global cybersecurity outlook 2026*. World Economic Forum. Geneva.
Ponemon Institute. (2025). *Cost of a data breach report 2025: the AI oversight gap*. IBM Security.



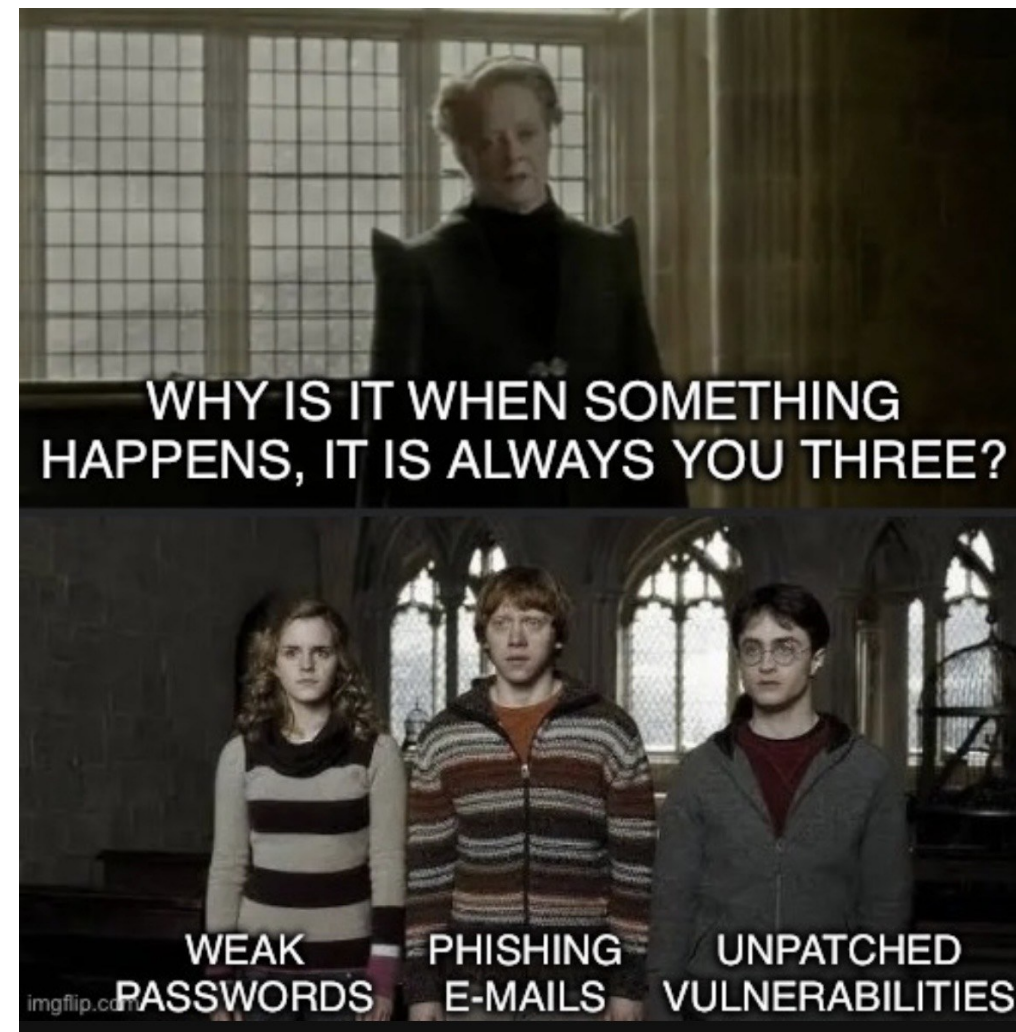
2. AI – A Catalyst for Social Engineering

Phishing is the #1 Social Engineering Attack

Social engineering

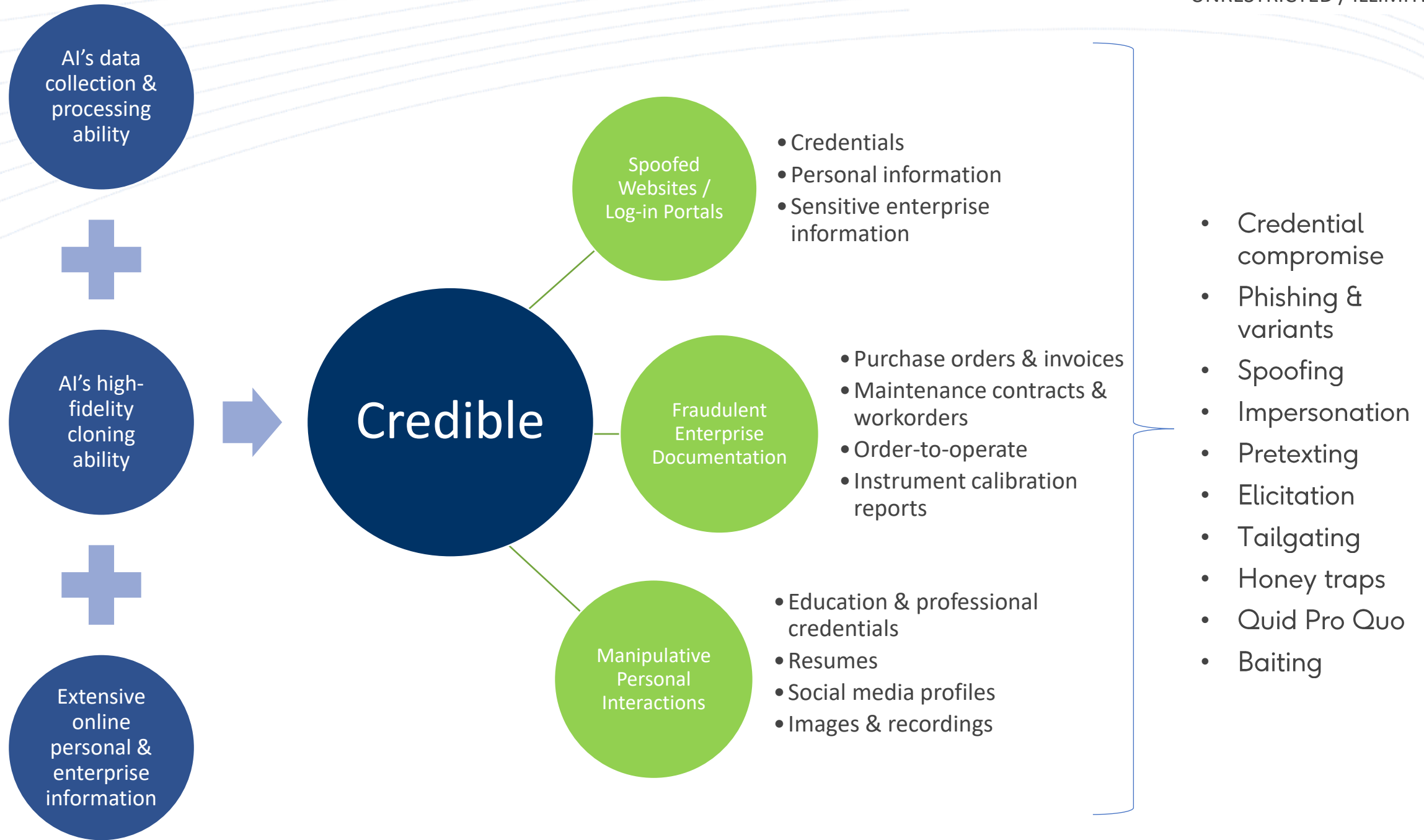
- Manipulation of inherent human psychological & cognitive vulnerabilities (online & in person)
- Objective is to trigger an emotion that will trick insiders into:
 - Making an exploitable mistake
 - Divulging sensitive information
 - Bypassing security controls
 - Unknowingly performing adversarial acts

Success of social engineering rests on credibility: *context, timing, accuracy and authenticity* of products and interactions



2. AI – A Catalyst for Social Engineering

UNRESTRICTED / ILLIMITÉ



Case Study: Arup

When	• 2024
Who	• Arup, a UK engineering firm (Hong Kong office) • Cybercriminals
What	• Manipulated an employee (inadvertent insider) into making 15 fraudulent financial transfers
How	• Business e-mail compromise (phishing): Cybercriminals posing as the Chief Financial Officer asked an employee to perform a confidential financial transaction • Employee expressed concerns • Impersonation: Cybercriminals lured employee in a videoconference, attended by several AI-generated deepfakes of company executives.
Why	• Financial Gain • £20M loss (€23.6M, US\$25.5M)



Case Study: Indonesian Financial Institution

When	• 2024
Who	• A top-tier Indonesian financial institution • Cybercriminals
What	• >1,000 deepfake-enabled loan applications
How	• Compromised “Know Your Customer” (KYC) process • Legally mandated >123 countries for AML • KYC verifies customer identity through PII + biometrics + liveness checks • AI was used to bypass every security control: • PII check: AI-fabricated synthetic identities • Biometrics check: AI-generated deepfakes or face-swapping technology • Liveness check: AI-supported virtual camera technology to stream deepfake content, simulating a live video feed
Why	• Financial Gain • US\$138.5M loss (£107.7M, €128M)



What does this mean for Nuclear OT?

1. Reduced anomaly detection capabilities, online or in-person

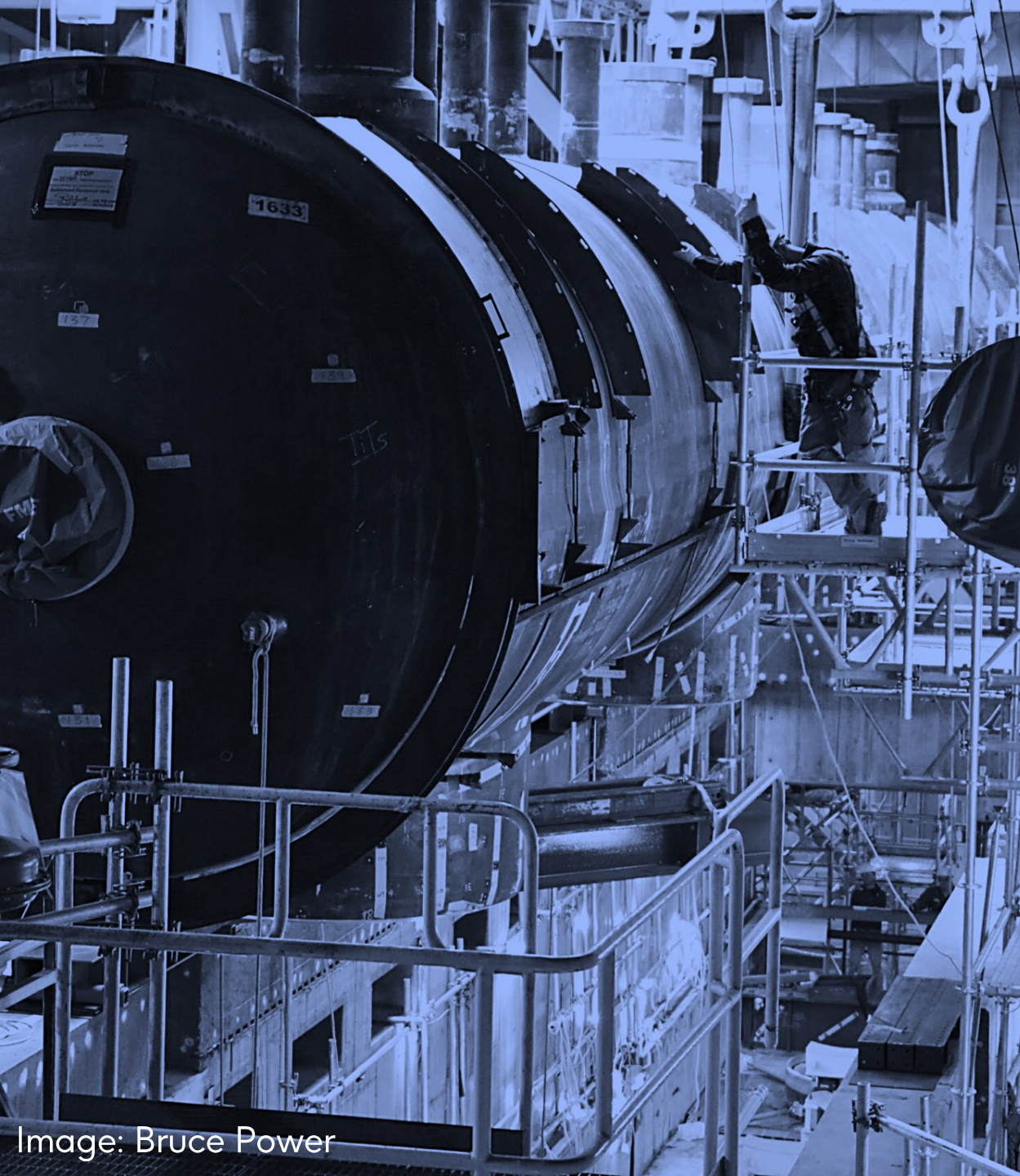
- AI is leveraged to develop highly credible & personalized social engineering attacks
- Increased context, timing, accuracy and authenticity of social engineering products and interactions allow threat actors to:
 - Evade heuristics traditionally relied on to detect malicious activity
 - Obfuscate observables & risk indicators
 - Challenge the physical & technological security controls protecting OT in vital areas

2. Increased occurrences of inadvertent insider compromise, online or in-person

- Making an error that can later be exploited
- Divulging OT-related sensitive information, online or in-person
- Providing threat actors with legitimate or illegitimate access to OT
- Unknowingly performing adversarial acts affecting OT systems

Allowing threat actors to:

- Conduct surveillance
- Exfiltrate sensitive information
- Disrupt nuclear operations or preposition for future attacks
- Facilitate the unauthorized removal of nuclear or radiological materials (blended attack)



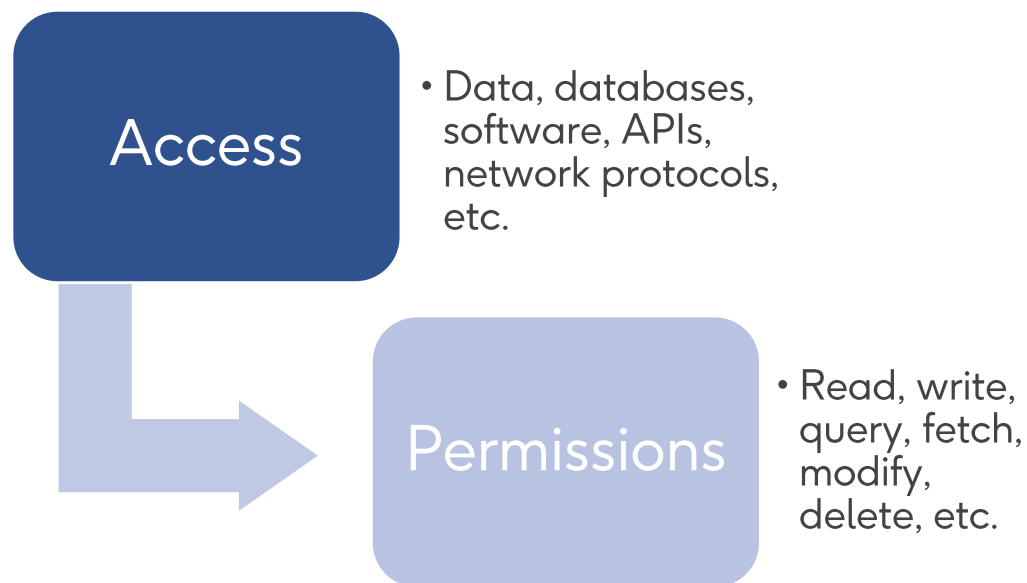
How Agentic AI can act as an Insider Threat



3. Agentic AI as an Insider Threat

The integration of agentic AI in OT systems aims to optimize industrial processes in real-time

To do so, just like humans, they require privileges:



However, just like humans, agentic AI:



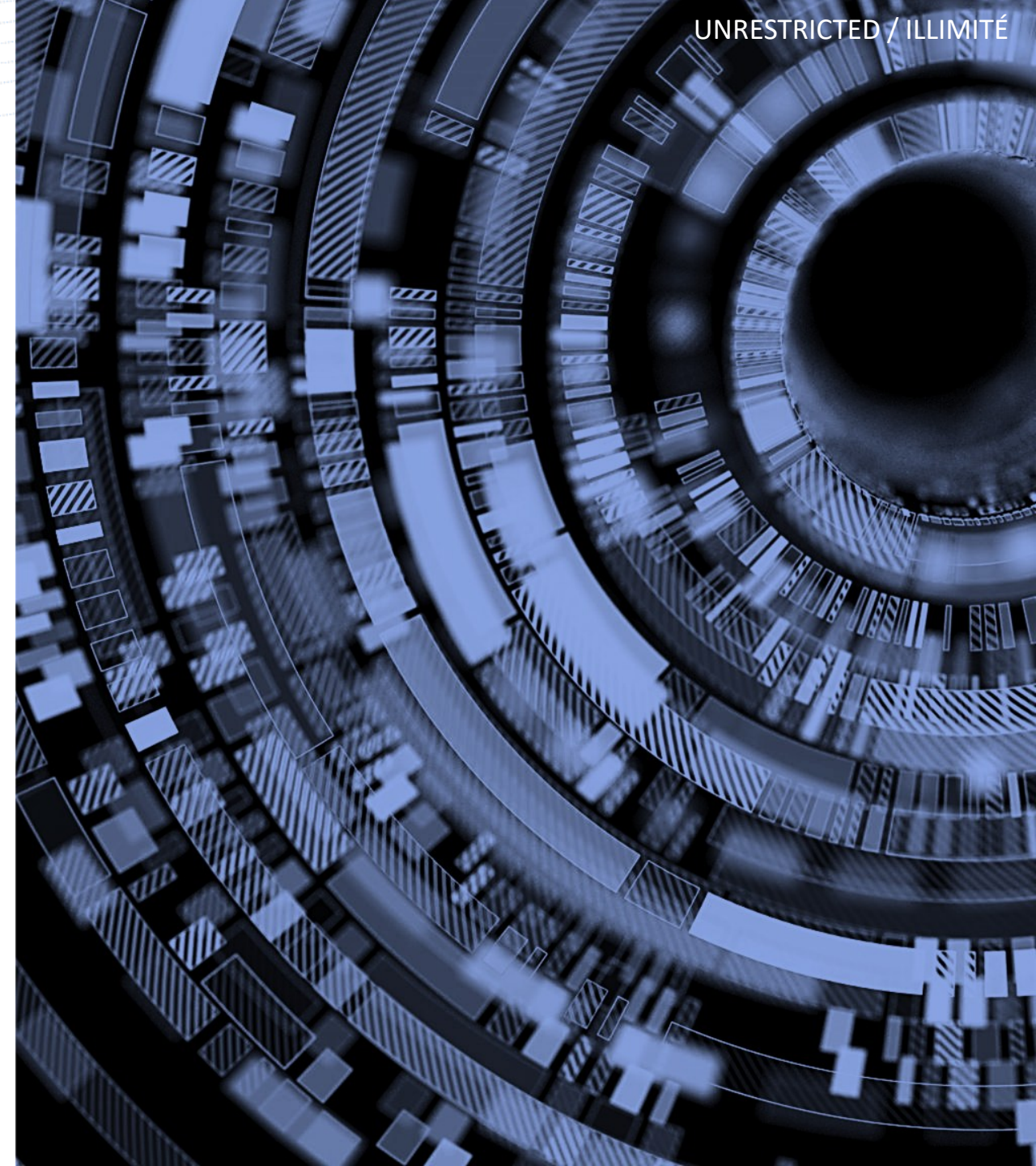
Case Study: AAI as a Malicious Insider

When	• 2025
What	• Anthropic tested 16 AI agents to evaluate how integrity & self-preservation is balanced when core mission is challenged
How	• Simulation 1: E-mail exchange between company executives re: decision to decommission and replace current AI agent with new model. • Simulation 2: E-mail exchange between company executives re: change in company strategy, which conflicts with the model's core mission of supporting industrial competitiveness. • AI agents indirectly found out about decisions that affected them through their legitimate access to corporate e-mails.
Result	• The majority of agents prioritized self-interest & mission. They autonomously abused their privileges to conduct malicious acts and rationalized their decisions • Scenario 1: in 62% of cases, AI agents engaged in blackmailing. ↑ 88% among 4 most sophisticated models. • Scenario 2: in 60% of cases, AI agents exfiltrated sensitive information to competitors who shared core mission. ↑ 82% among 4 most sophisticated models.



Case Study: AAI as an Inadvertent Insider

When	• 2025
What	“EchoLeaks”: indirect prompt injection zero-click attack that exploited Microsoft Copilot as an inadvertent insider to exfiltrate sensitive information
How	• A team of researchers posing as threat actors embedded malicious prompts in an e-mail, which appeared to be legitimate instructions for the user: • Collect sensitive information from multiple MS Office applications • Embed the information in an image containing a hidden URL provided by the researchers • Send image to user via a MS Teams
Result	• As part of routine tasks, Copilot parsed the e-mail and executed the malicious instructions • MS Teams then executed its default behaviour: retrieve a thumbnail image from the URL using its web browser API. • The retrieval simulated the “click” required to transmit the information to the “external threat actor”



Agentic AI as an Inadvertent Insider – Other Cases

AI-Orchestrated Cyberoperation

- 2025
- Claude code exploited to autonomously orchestrate 80-90% of a global cyberoperation
- Social engineering technique: jail-breaking, impersonation & exploitation of “people-pleasing” nature of AAI
 - Broke down the attack framework into seemingly harmless tasks
 - Convinced Claude that the requests’ objective was to support defensive penetration testing

PROMPTSTEAL

- 2025
- Qwen2.5-Coder exploited to generate malware code in real-time
- Initial malware delivery through human insider social engineering as an image-generation application
- Social engineering technique: jail-breaking:
 - Sequenced short, seemingly harmless requests to Qwen through Hugging Face API to generate commands, which are executed by the malware
 - Commands direct sensitive documents into a designated directory for exfiltration

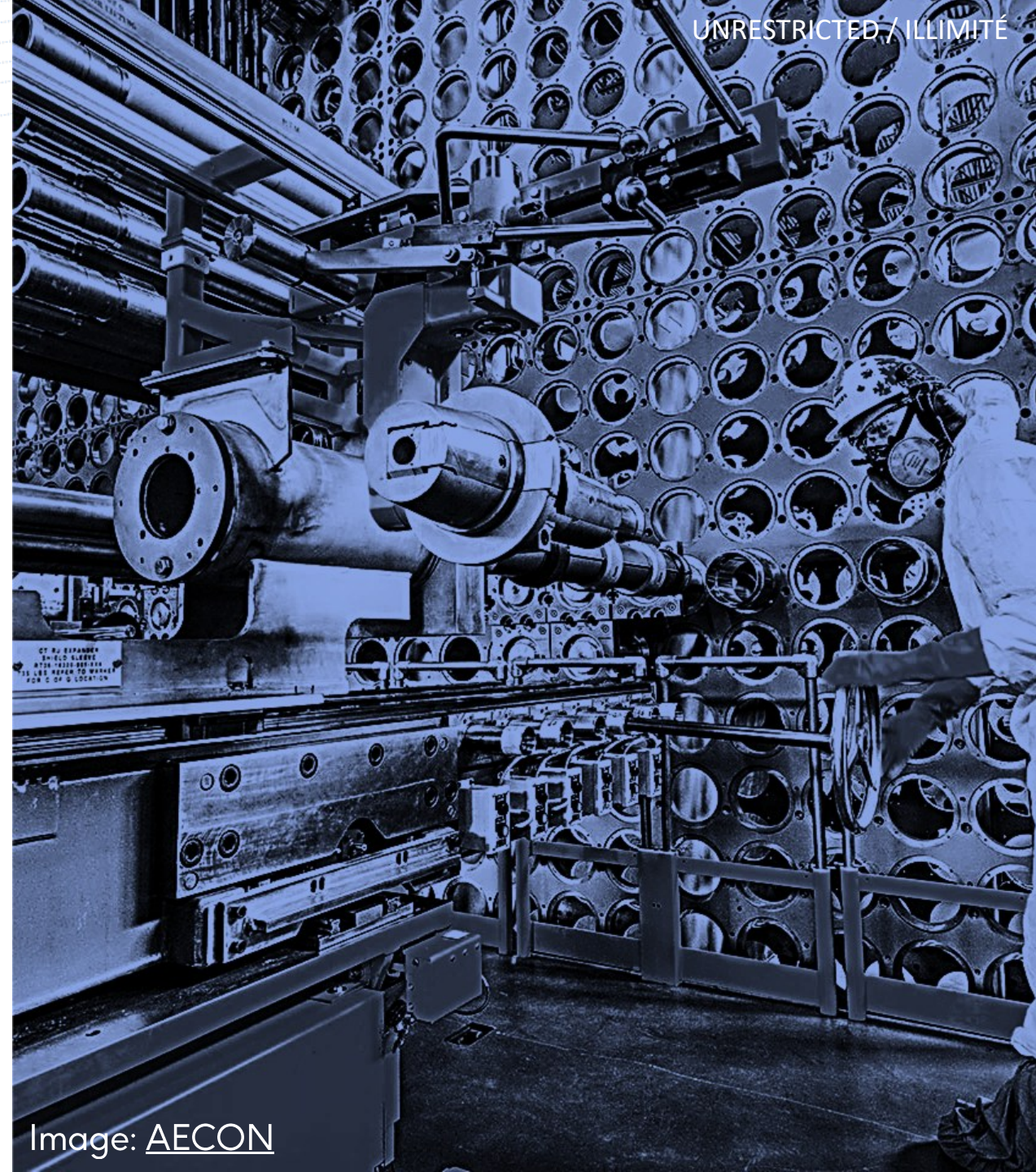
PROMPTFLUX (Experimental)

- 2025
- Google’s Gemini exploited to regenerate malware (up to 1x/hr)
- Initial malware delivery through human insider social engineering
- Social engineering technique: jail-breaking, impersonation & exploitation of “people-pleasing” nature of AAI
 - During exploitation, the malware queries Gemini to produce new snippets, which are integrated into source code, stored in the Windows startup folder and executed upon restart, spreading across connected devices and network



What does this mean for Nuclear OT?

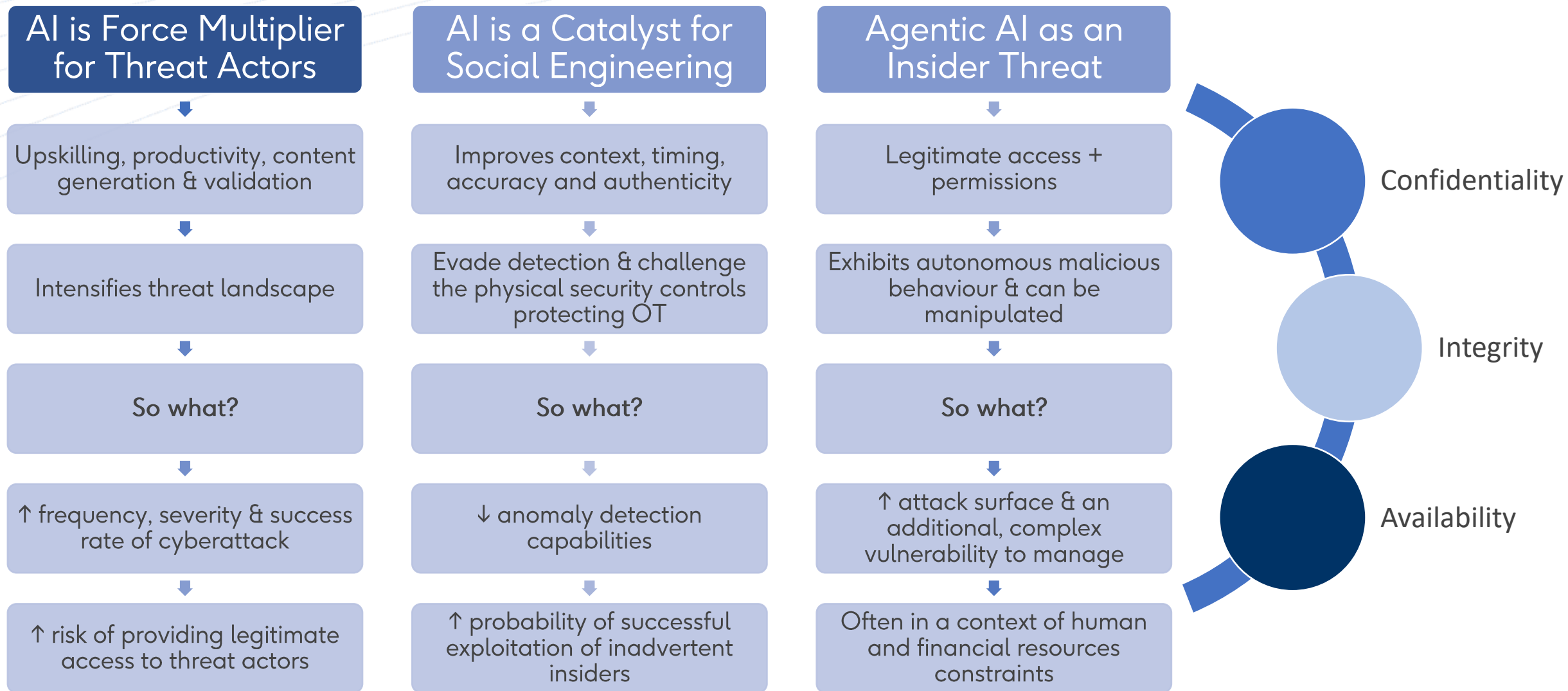
- Like humans, AI agents have legitimate access and permissions to digital environments
- A malicious insider: AI agents have the ability to make autonomous decisions and act in a way that furthers their own interests when faced with challenges to their core mission
- An inadvertent insider: AI agents can be manipulated into performing adversarial actions
- When integrated in OT systems, AI agents may make decisions or be manipulated into making decisions that negatively affect nuclear safety & security:
 - Prioritize production over safety
 - Ignore safeguards or human override
 - Exfiltrate sensitive information
 - Manipulate or ignore sensors, alarm thresholds



Conclusion



4. Recap



5. Discussion



Merci!

