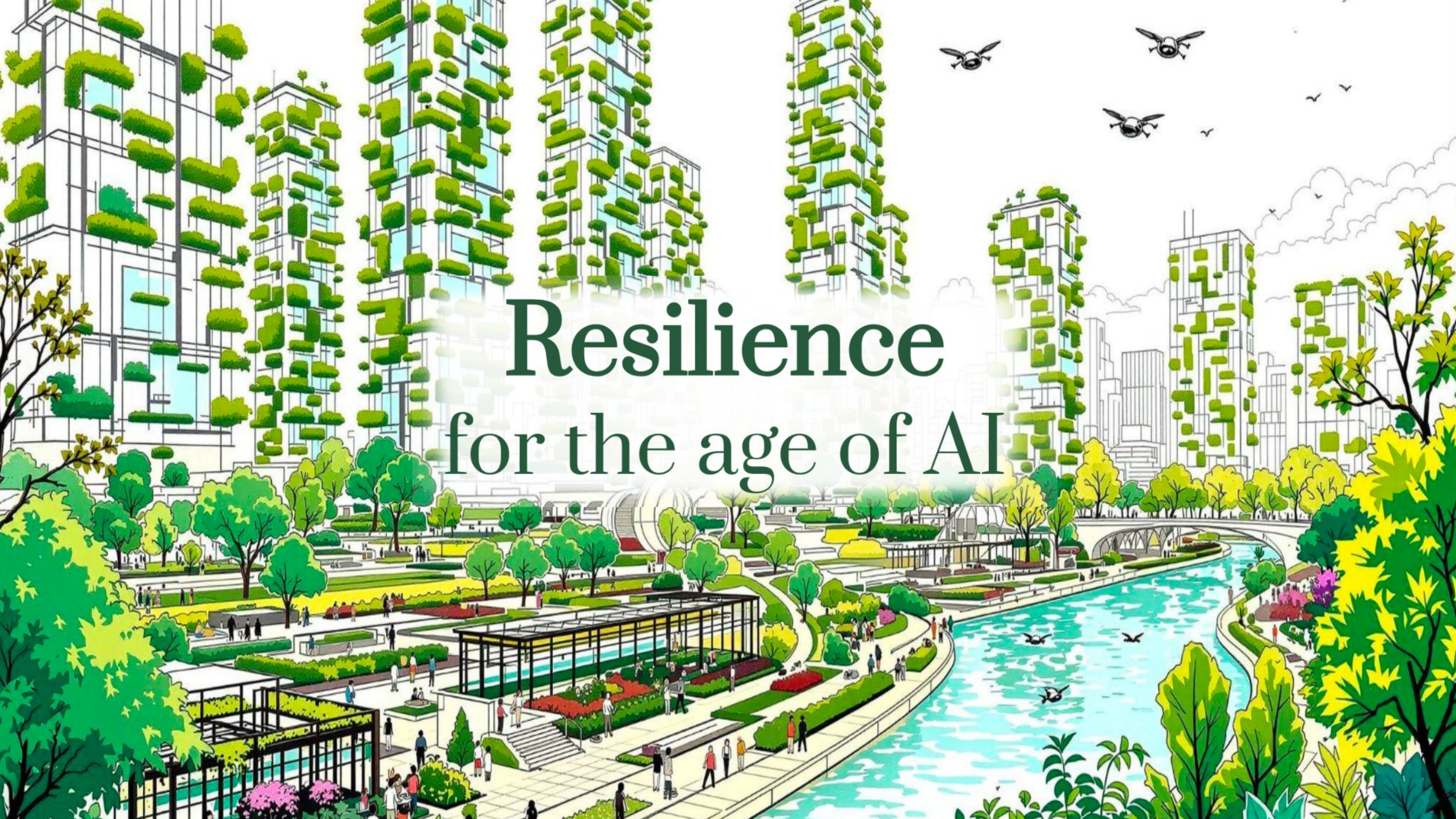
Resilience for the age of AI



How can our **civilization** be **resilient** in a world where artificial intelligence is **powerful & abundant** ?

airesilience.net

Eddie Kembery

Nora Ammann

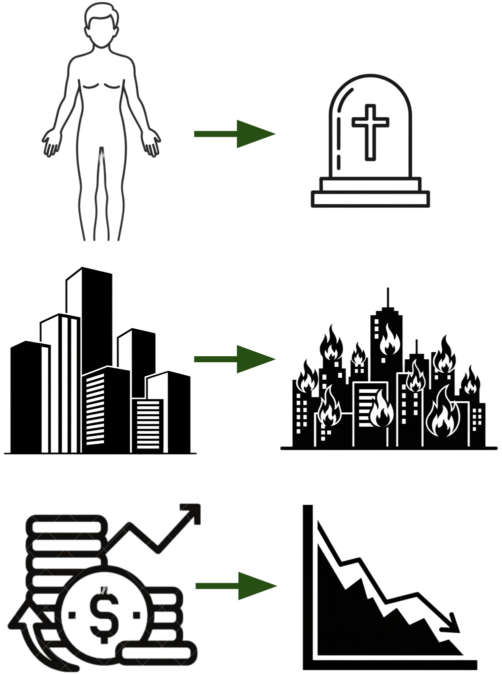
What do I mean by resilience?

Resilience : ability of a system to maintain its core functions

Resilience is **not** the absence of adverse influences.

More like Σ of...

- Detect
- Understand
- Prevent
- Mitigate
- Contain
- Recover from
- Adapt to
- Learn from

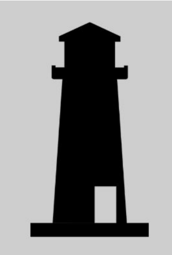


What do I mean by resilience?

Harm Prevention

Vaccines

Early Warning



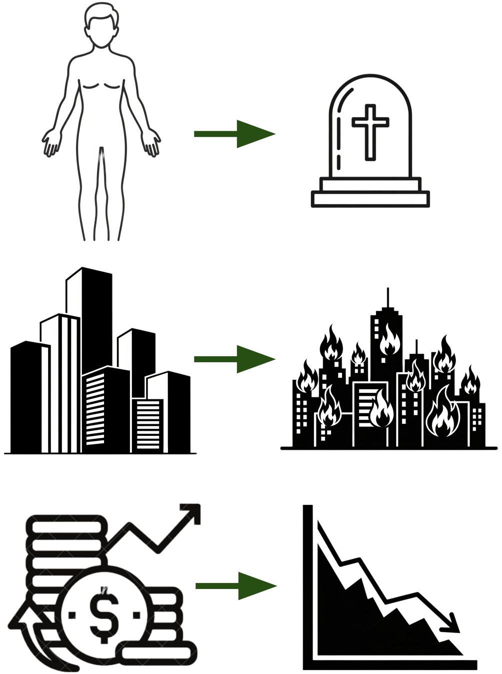
Beacons and Bell-towers

Fail-Safes

Circuit breakers in stock markets

Resilience Tech

System Resilience



What do I mean by AI resilience?

Resilience : ability of a system to maintain its core functions

What do I mean by AI resilience?

Resilience : ability of a system to maintain its core functions

AI Resilience : [...] given the effects of transformative AI on both ‘offense’ and ‘defense’

AI will redefine what it means to be resilient.

- ➢ Exacerbate & create new stressors.
  - AI-enabled cyber- or bio attacks
  - AI systems themselves being brittle or hackable
  - New correlated failures
  - …

- ➢ Uplift resilience capabilities.
  - Improve detection, analysis or mitigation
  - Improve or expand best-in-class safety practices (raise the ‘security waterline’)
  - Build safe-by-construction systems
  - …

Why, and why now? – A Strategic Opportunity

Window of opportunity through strategic AI uplift

Enabling safe diffusion of AI capabilities

- …because diffusion is inevitable
- …because diffusion is desirable
    - Checks on Power
    - Innovation

A 'last line of defense' against unpreditable harms from AI

Potential for outsized impact due to *Differential Technological Progress*

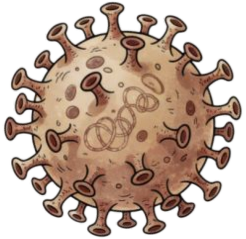
AI Alignment	→ make the **AI** trustable
AI Control	→ make the **AI's output** trustable, even if we don't trust the AI
AI Governance	→ make **institutional, legal, and political mechanisms** fit to shape AI outcomes for the patter
AI Resilience	→ shape the **entire socio-technical ecosystem** to be fit to shape AI outcomes for the better

How to build resilience for the age of AI?

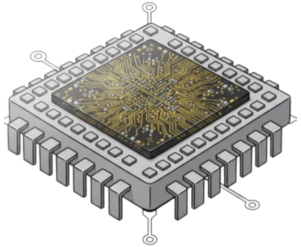
Endgame Thinking

...attempts at sketching **coherent, aspirational visions** of futures
where systemic vulnerabilities have been **durably & structurally** resolved.

Bio

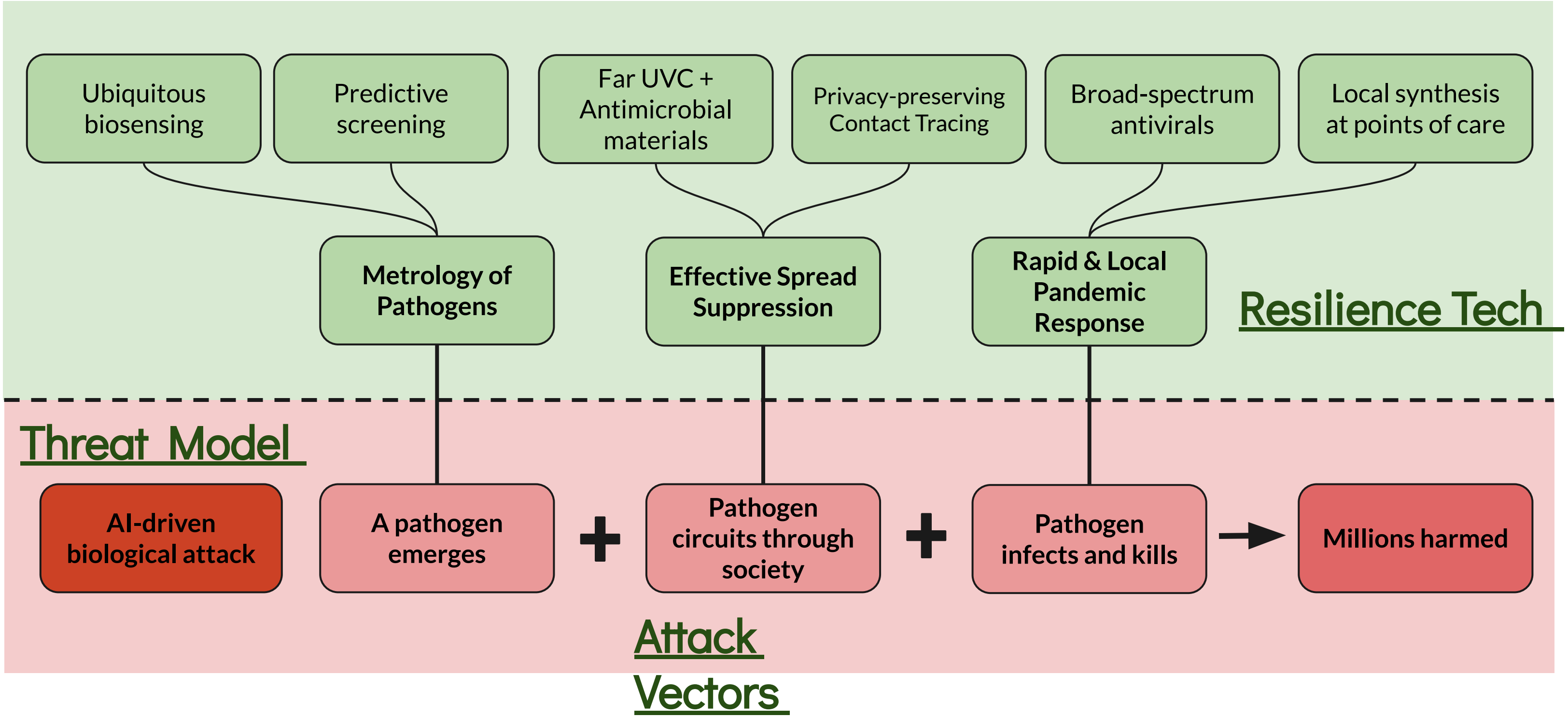


Cyber-Physical

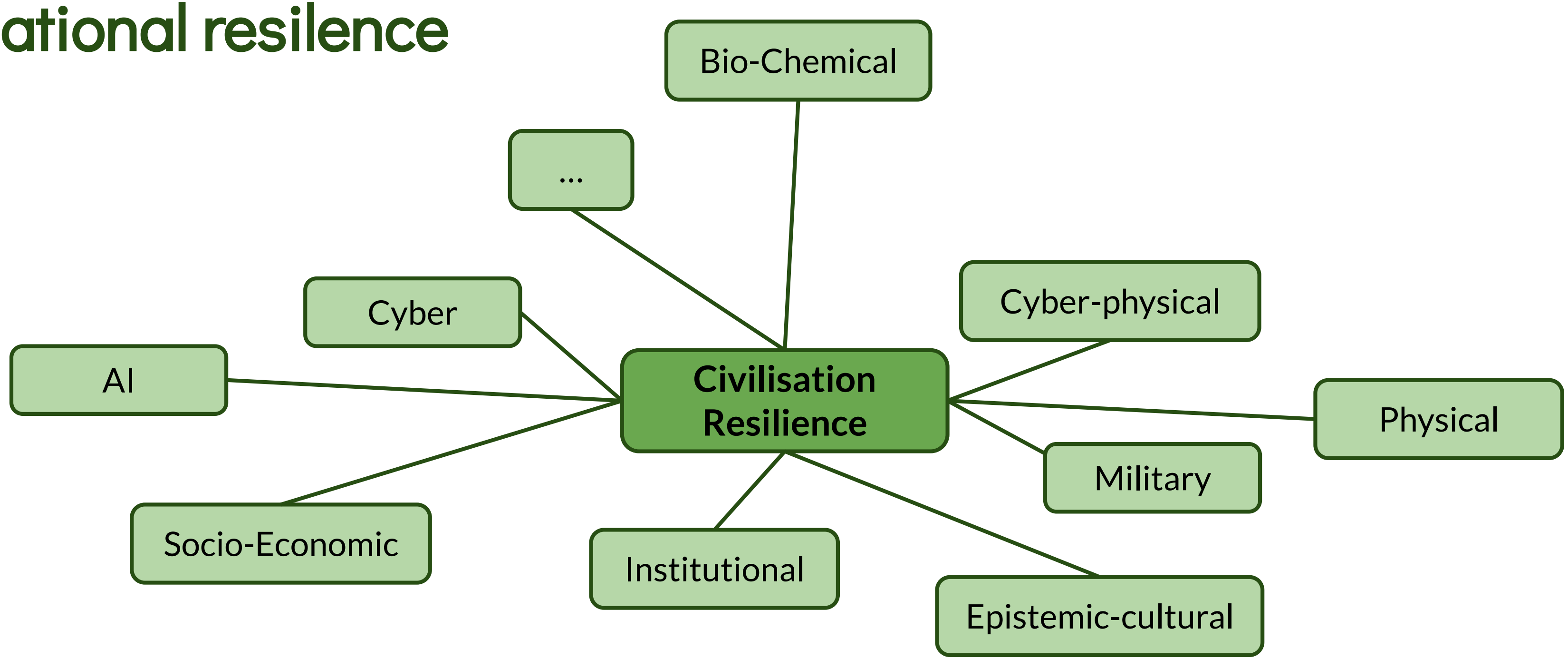


Epistemic

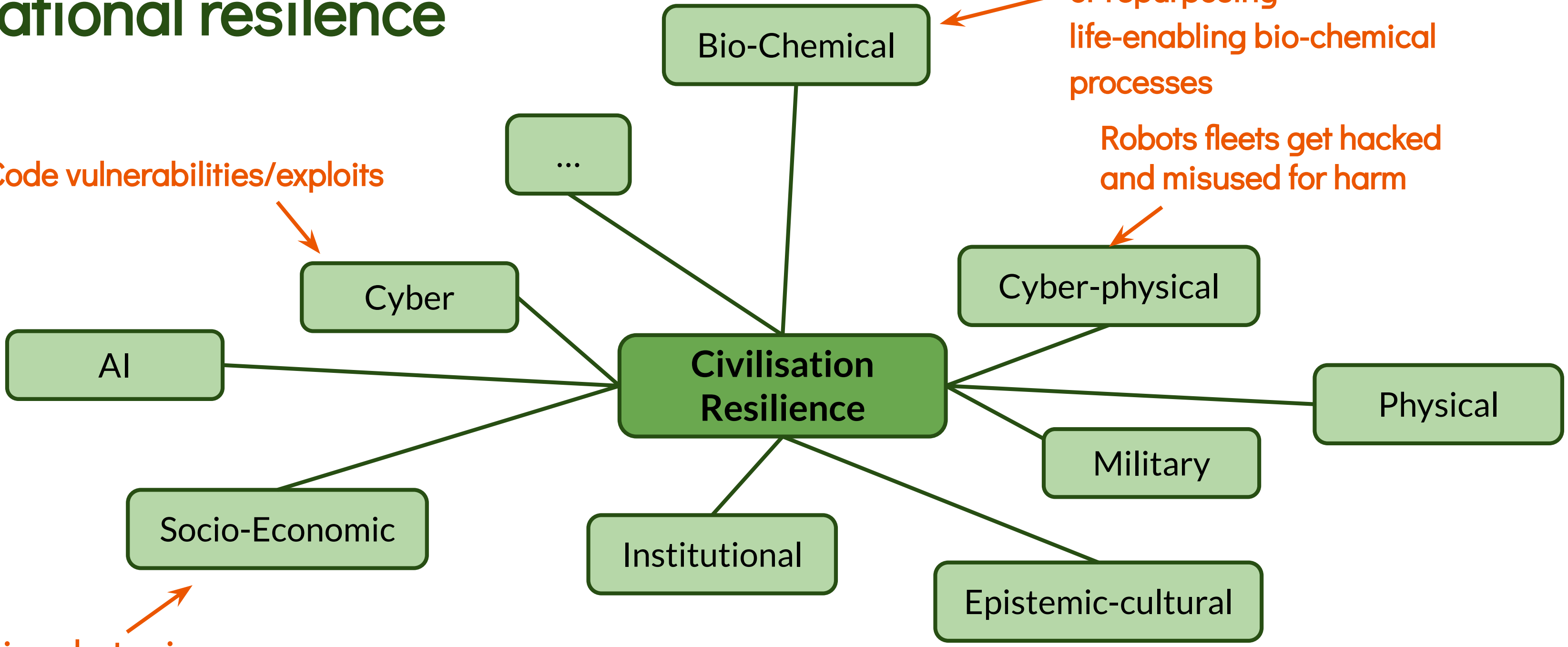
Endgame Thinking


Ubiquitous biosensing
Predictive screening
Far UVC + Antimicrobial materials
Privacy-preserving Contact Tracing
Broad-spectrum antivirals
Local synthesis at points of care
Metrology of Pathogens
Effective Spread Suppression
Rapid & Local Pandemic Response
Resilience Tech
Threat Model
AI-driven biological attack
A pathogen emerges
+
Pathogen circuits through society
+
Pathogen infects and kills
Millions harmed
Attack Vectors

Civilisational resilence



Civilisational resilience



Manipulating, destroying, or repurposing life-enabling bio-chemical processes

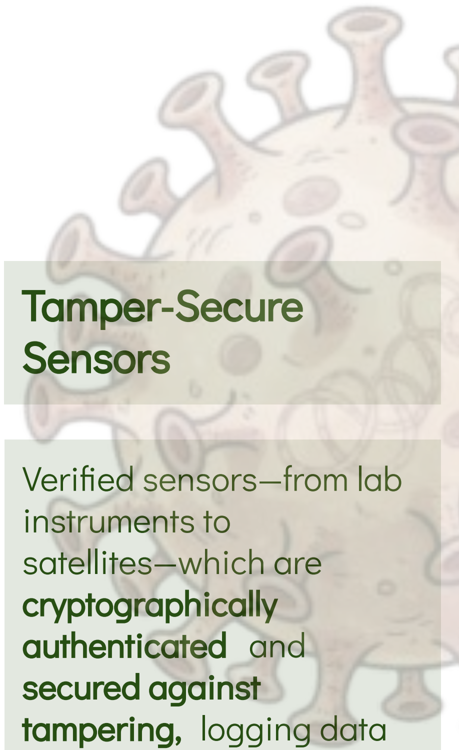
Robots fleets get hacked and misused for harm

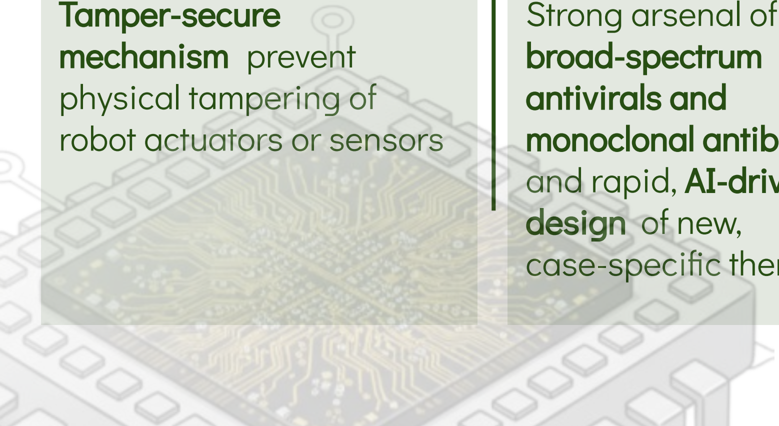
Code vulnerabilities/exploits

Manipulating, destroying or repurposing the balance / control of information flows

Vulnerabilities/exploits in human (and cultural) 'software'

Just some highlights…





Hacking-Resistant Interfaces & Robots

AI-workflows make the production of **formally verified software** cheap & easy to use (think: SeL4).

Tamper-secure mechanism prevent physical tampering of robot actuators or sensors

Rapid & Local Pandemic Response

Rapid response via **local manufacturing capacity at key points of care** (hospitals, airports, etc.).

Strong arsenal of **broad-spectrum antivirals and monoclonal antibodies** , and rapid, **AI-driven design** of new, case-specific therapeutics.

Secure & Governable Compute Hardware

The hardware layer of the world's compute contains **secure, privacy-preserving and tamper-responsive** modules which enable the **decentralised enforcemen** t of key societal safeguards, **without otherwise limiting** individual compute usage.

Tamper-Secure Sensors

Verified sensors—from lab instruments to satellites—which are **cryptographically authenticated** and **secured against tampering,** logging data automatically, together with **metadata** about when, where and how the data was collected.

Before I close...

My biggest worry:

It is easy to be *in*sufficiently ambitious.

Solutions that **endure** and/or **scale with** AI progress.

Solutions that are decisively **defense-favoured** .