Building the Immune System for a Hybrid World

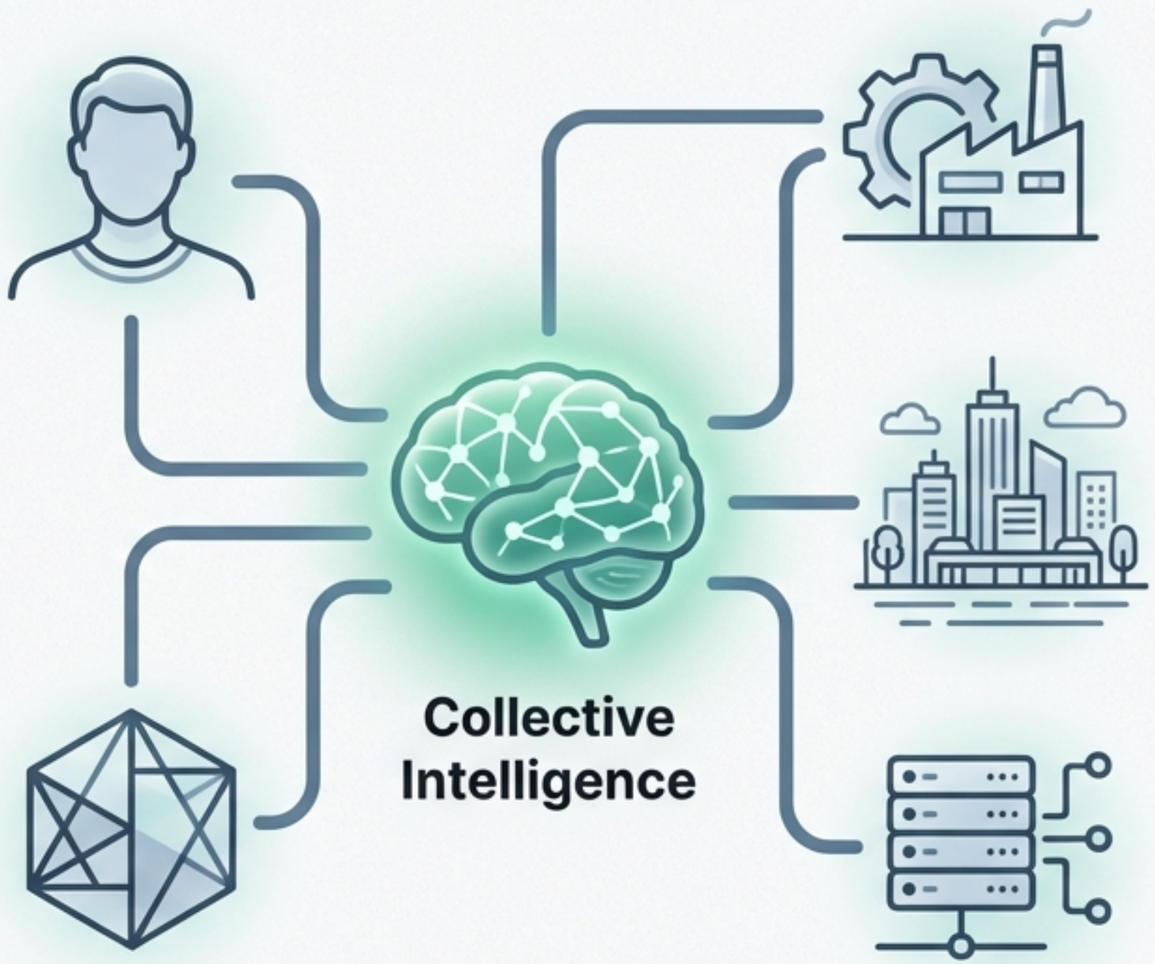
A New Paradigm for Resilience Against Cognitive Threats

Insights from the IMMUNE & WARN Projects

Our World is Becoming a Hybrid Society

We are building a new civilization based on collaborative intelligence, where humans and AI jointly manage our most critical infrastructure.

- **Industry 4.0/5.0**: The evolution from smart factories to human-machine symbiosis.
- **Cyber-Physical-Social Systems**: These are not just automated factories, but complex systems where humans remain the key decision-makers, supported by AI.
- **Collective Intelligence**: The core process is collaborative decision-making between people and autonomous AI agents. This integration is our greatest strength and a new source of vulnerability.

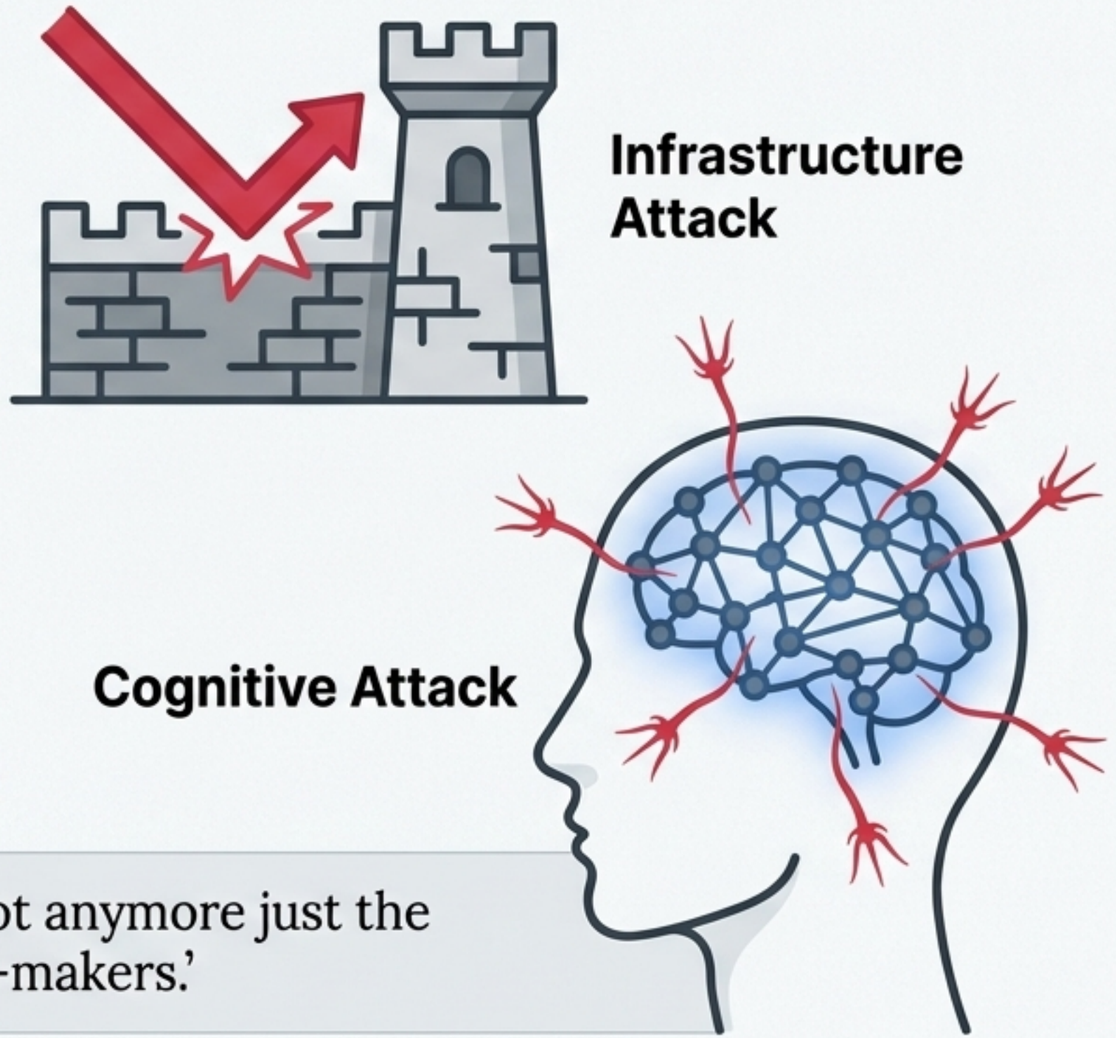


A New Threat Targets Our Minds, Not Our Machines

Key Concept: Hybrid Threats are coordinated activities targeting systemic vulnerabilities to influence decision-making. The goal is cognitive hacking.

Elaboration

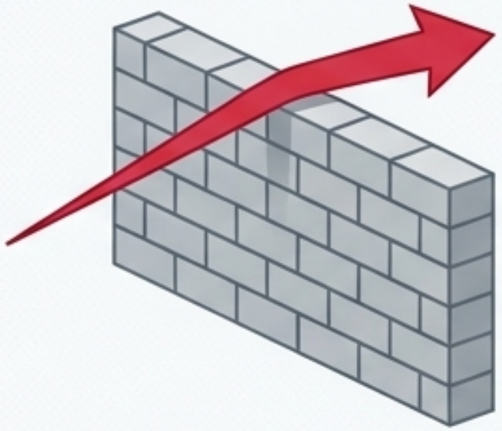
- **The Target:** The minds of decision-makers within the Hybrid Society.
- **The Goal:** To alter decisions to benefit the attacker's strategic goals.
- **The Nature of the Threat:** These attacks are stealthy and disguised, exploiting the cognitive vulnerabilities of both humans and AI.



'The targets for the external attacks nowadays are not anymore just the infrastructures but mainly the minds of the decision-makers.'

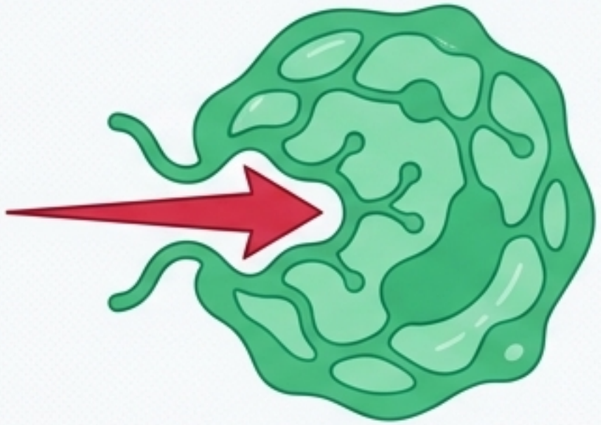
We Don't Need Higher Walls. We Need an Immune System.

Fortress Model (Old Paradigm)



- Static defenses
- Assumes predictable threats
- Protects infrastructure
- Fails when breached

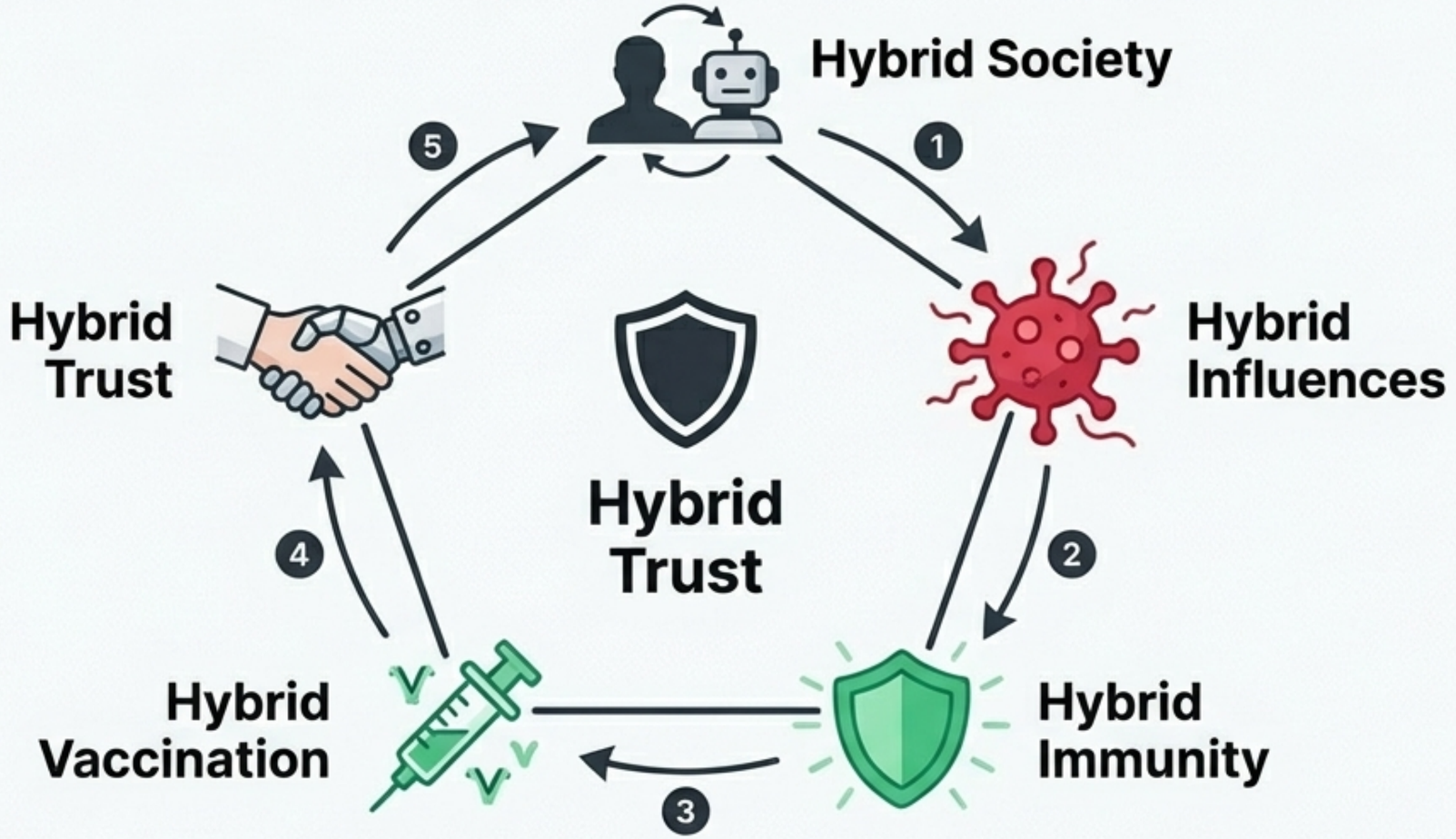
Immune System Model (New Paradigm)



- Adaptive defense
- Learns from exposure
- Protects cognitive processes
- Builds resilience

Traditional security is like building a fortress. For cognitive threats, we need a biological approach: a Hybrid Immunity capable of identifying, adapting to, and neutralizing threats targeting our decision-making processes.

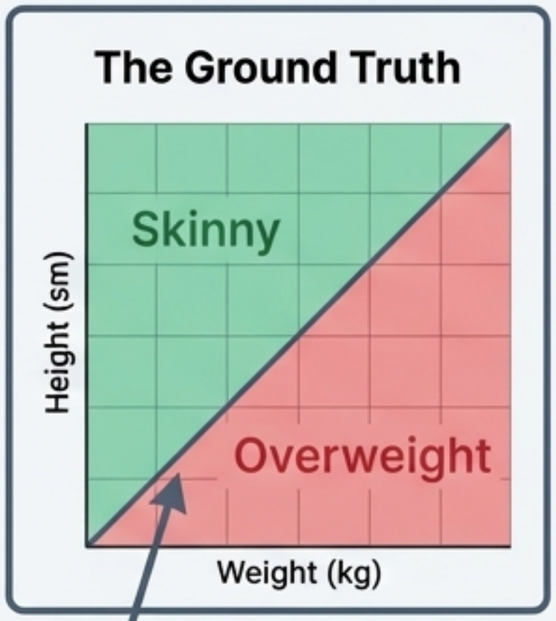
The Genetic Code of a Resilient Society: 5H4TRUST



This framework provides a roadmap for developing a hybrid trust in a hybrid society, driven by a hybrid vaccine that builds hybrid immunity against hybrid influences.

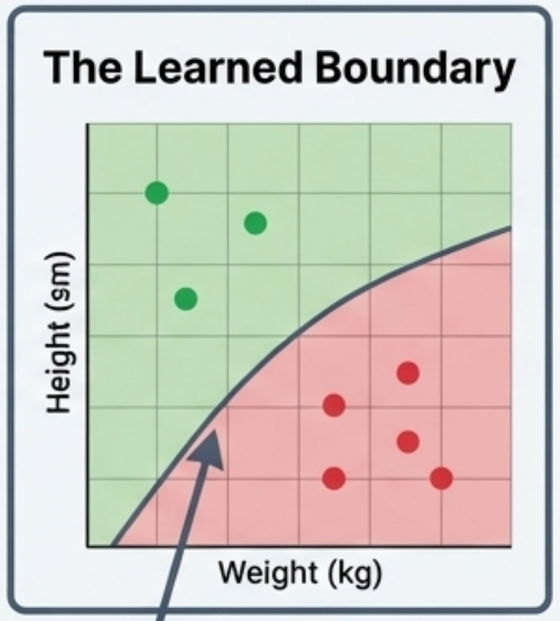
How Attacks Exploit the 'Grey Zones' in Decision-Making

We'll use a simple AI classifier to illustrate the core problem. The goal is to classify a person as "skinny" or "overweight" based on height and weight.

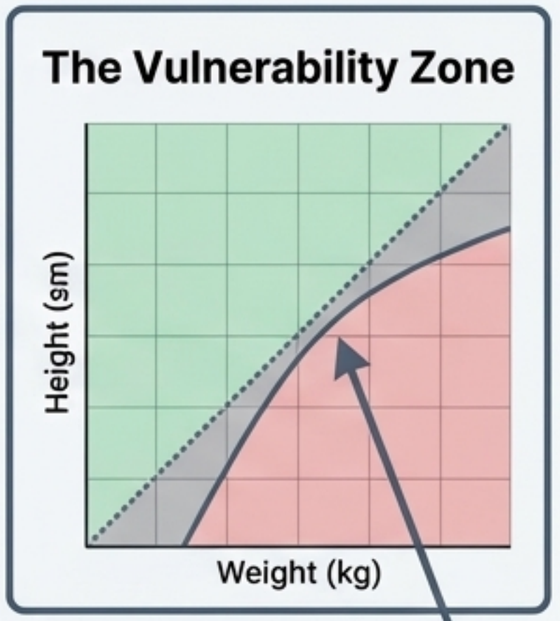


The 'hidden' correct decision boundary (the ground truth).

Limited training data points are given to the AI.



The decision boundary the AI *actually* learns.



Vulnerability Zone: An area of potentially incorrect decisions.

Any machine learning model has 'vulnerability zones'—areas where its learned understanding of the world is imperfect. Attackers exploit these zones.

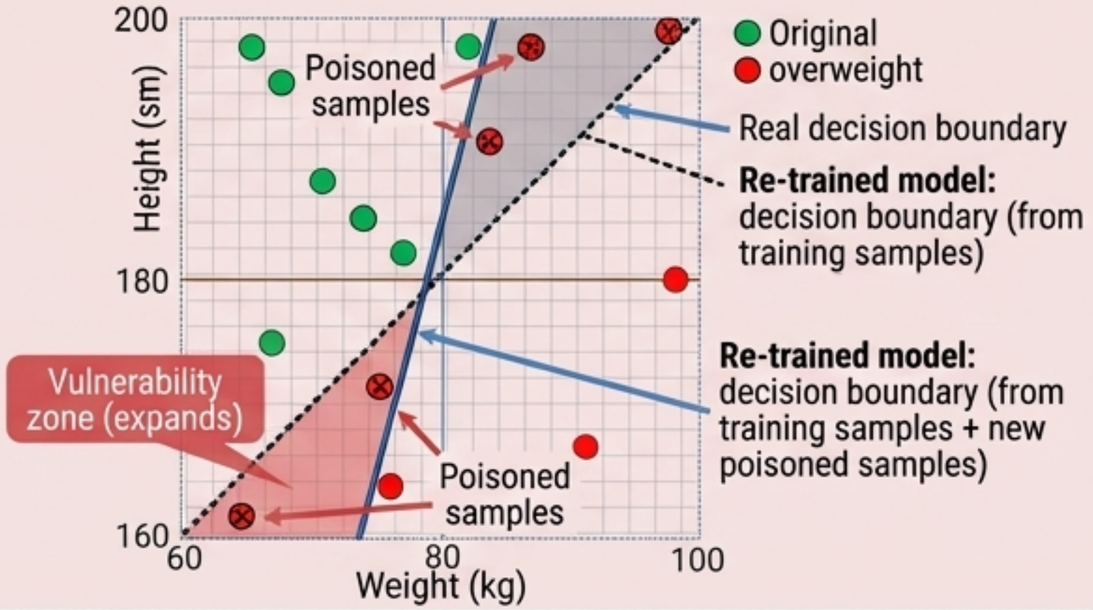
Poison or Vaccine? The Difference is Intent.

The same technique used to attack a system can be used to strengthen it.
The key is how we label the adversarial data we introduce.

Poisoning: Expanding Vulnerability

Adversarial samples are generated in the vulnerability zone and given ***incorrect*** labels.

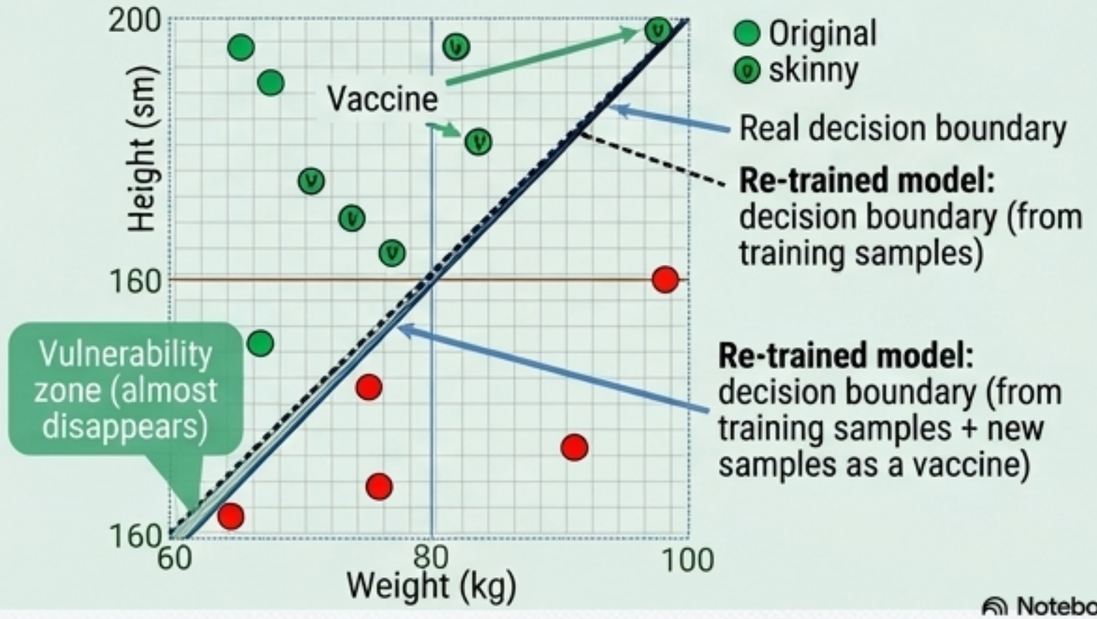
The AI **retrains** on this "poisoned" data, and its learned decision boundary **moves *further*** from the truth, expanding the vulnerability zone.



Vaccination: Building Immunity

The ***same*** adversarial samples are generated but are given the ***correct*** labels.

The AI **retrains** on this "vaccine" data, and its boundary **moves *closer*** to the truth, shrinking or eliminating the vulnerability zone.



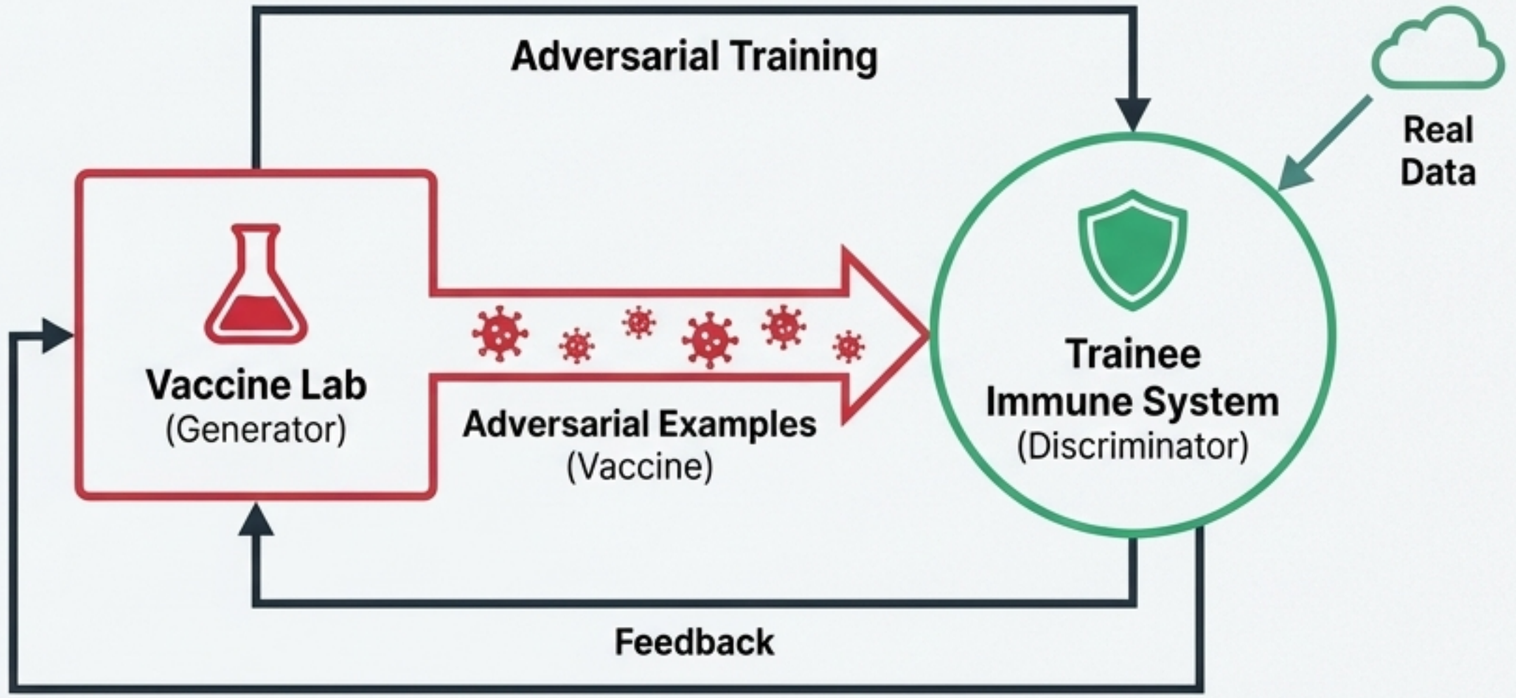
Clinical Trial I: Vaccinating Artificial Intelligence (The IMMUNE Project)

Objective

To build “digital immunity” for AI systems to make them robust against adversarial attacks like data poisoning.

The Core Technology

Modified Generative Adversarial Networks (GANs).

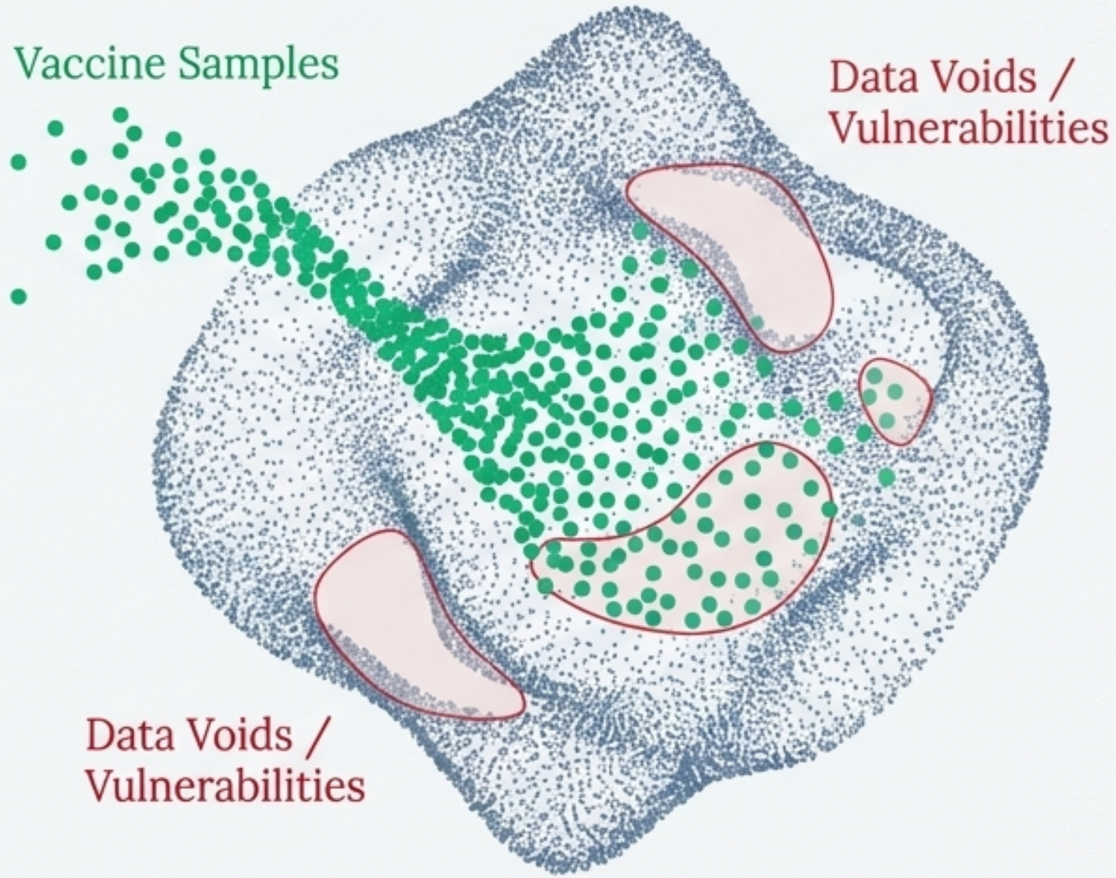


The **Vaccine Creator** makes sophisticated adversarial examples (the “vaccine”) to challenge the other network. The **Immune System** learns to distinguish real data from the “vaccine” data, making its own decision boundaries more robust. The two train against each other, effectively “vaccinating” the system against future attacks.

Key Findings from the AI Immunity Trials

The IMMUNE project developed several practical techniques for building resilient AI, proving the effectiveness of the ‘digital vaccine’ approach.

- **Data Voids as Vulnerabilities**: Discovered that ‘voids’ or gaps in training data manifolds correspond to potential vulnerabilities that can be exploited.
- **Vaccination by Filling Voids**: Showed that smartly filling these voids with correctly-labeled adversarial samples acts as a powerful digital vaccine.
- **Reliable Decision Boundaries**: Proved that adversarial training helps build well-formed, reliable decision boundaries, reducing the ‘grey zones’ where an AI can be manipulated.
- **Digital Cloning for Defense**: The same adversarial learning process used for immunity can also learning process used for immunity can also create ‘digital cognitive clones’ of human decision-making, which can be used as a core defense component.



The Immunity Principle Extends from AI to Humans

Key Insight: While the “hardware” is different (silicon vs. biological neurons), the abstract mechanism of decision-making has similar vulnerabilities to adversarial influences.
The Connection: The lessons learned from vaccinating AI in the IMMUNE project can be adapted to build “cognitive immunity” in human decision-makers.
Next Step: This is the core objective of the WARN Project (“Academic Response to Hybrid Threats”): to develop a “cognitive vaccine” for humans through specialized training.

Clinical Trial II: A ‘Cognitive Vaccine’ for University Students (The WARN Project)

The Vaccine

Specially designed university courses using adversarial training in the form of structured disputes.

The Subject Matter

Training focuses on cognitively vulnerable areas—dilemmas where society lacks a shared opinion, making individuals susceptible to manipulation.

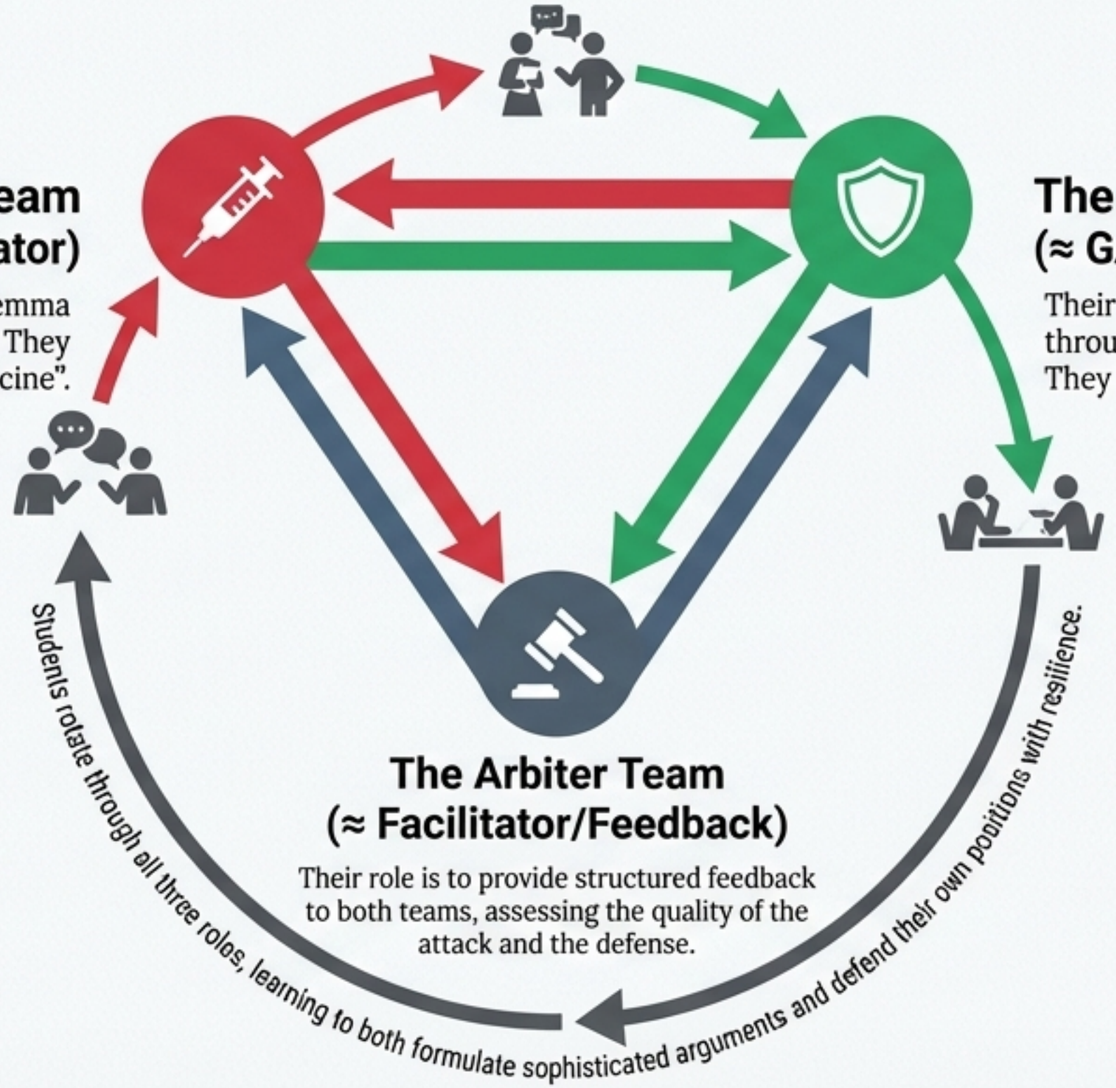
The Method

Before training, each student performs a cognitive self-assessment, rating the personal importance of each issue and their confidence in their ‘YES/NO’ stance.

Example Dilemma-Issues

The Human Training Ground Mirrors the AI Architecture

The classroom dispute simulates the adversarial process of a GAN.



The Attacker Team (≈ GAN Generator)

Their goal is to create a compelling dilemma that causes maximum confusion. They create the "vaccine".

The Defender Team (≈ GAN Discriminator)

Their goal is to analyze the dilemma, reason through it, and make a clear, confident choice. They are the "immune system" being trained.

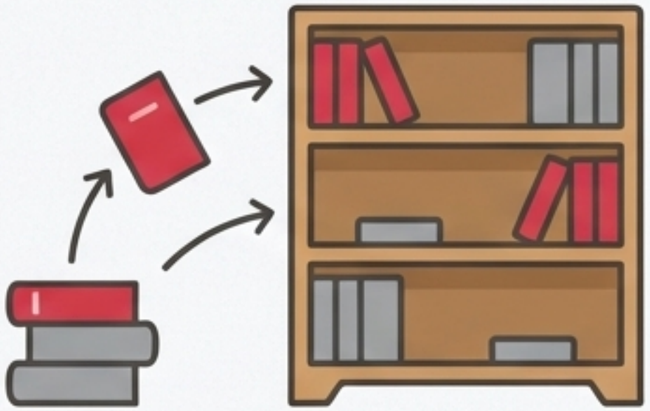
The Arbiter Team (≈ Facilitator/Feedback)

Their role is to provide structured feedback to both teams, assessing the quality of the attack and the defense.

The Goal Isn’t to Change Minds. It’s to Make Them Stronger.

A Radical Shift in Learning Objectives

The Old Goal of Education



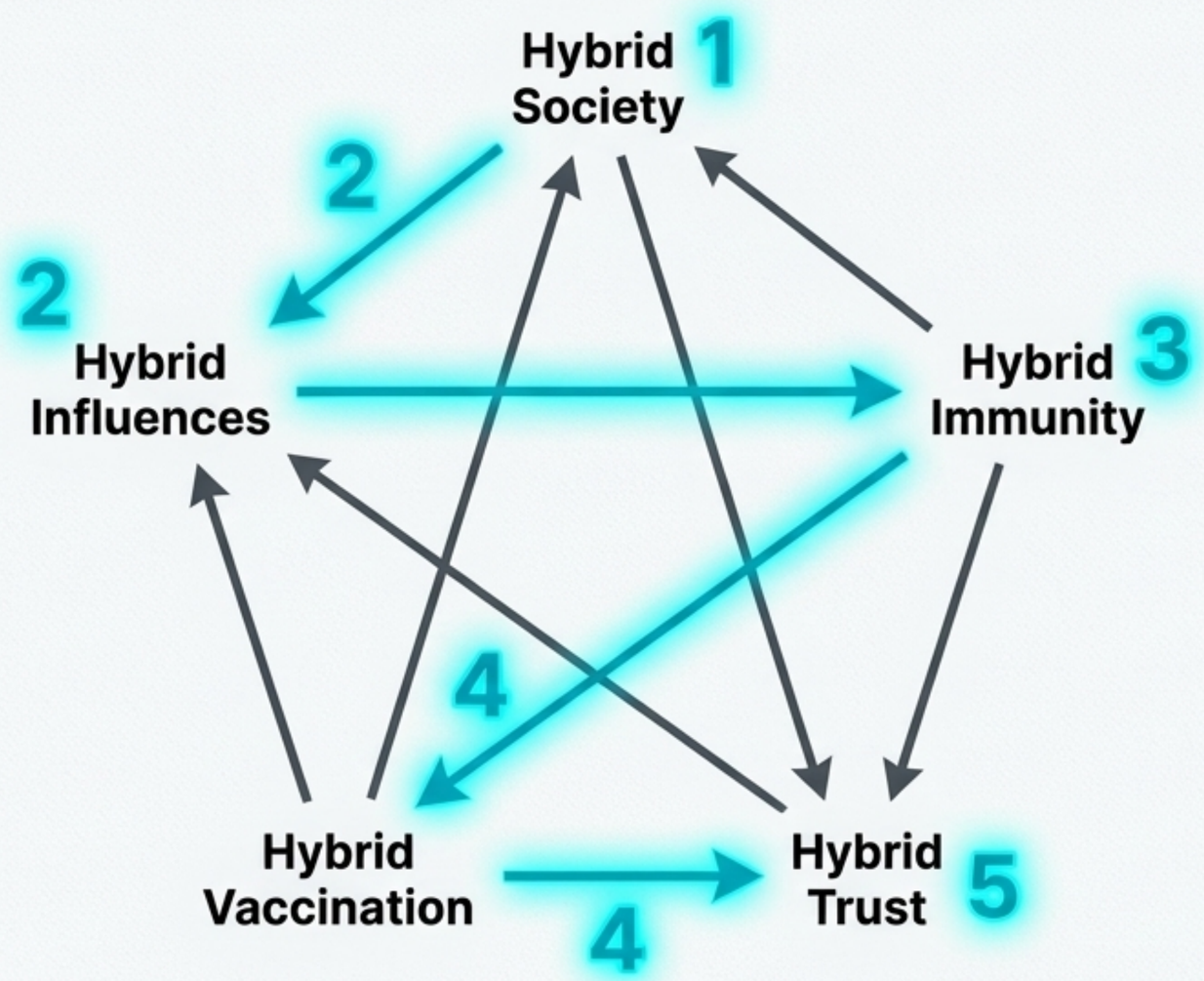
To *update* or add to a student’s personal knowledge, skills, beliefs, or values.

The New WARN Objective

To ensure the *robustness and resilience* of the knowledge, skills, beliefs, and values a student *already possesses.*

‘The new WARN adversarial training approach will be as different from zombifying students with new content as the process of vaccination is different from the process of poisoning.’

From Vaccination to Hybrid Trust



1. In our **Hybrid Society**...
2. ...we face **Hybrid Influences** that target cognition.
3. We must build **Hybrid Immunity** to defend ourselves.
4. This immunity is trained through **Hybrid Vaccination** (for both AI and humans).
5. Leading to the ultimate goal: **Hybrid Trust**, where human and digital partners can collaborate effectively and securely, resilient to manipulation.

Trust is the cornerstone of any collaborative system. In a hybrid world, trust itself must be hybrid—and actively defended.

The future of security is not architectural, it is biological. It will be won not by building higher walls, but by cultivating deeper resilience.