AI GOVERNANCE SHOULD PRIORITIZE CONTROL AND KNOWLEDGE BOUNDARIES OVER LIMITING INTELLIGENCE

Hendrik Luuk

Machine Learning and Computational Sciences, Pfizer Research and Development, Pfizer Inc., Cambridge, MA, USA

Department of Physiology, University of Tartu, Tartu, Estonia

arXiv paper

Pfizer Outdo Yesterday™

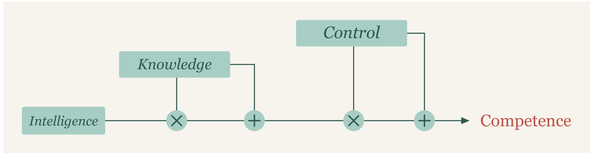
ABSTRACT

The rapid proliferation of autonomously acting AI agents makes the governance of agentic systems urgent. We introduce a comparative definition of epistemic intelligence that separates hypothesis-selection efficiency from available knowledge. This distinction motivates the competence circuit (IKC + KC + C), which relates intelligence, knowledge, and control to real-world impact. Motivated by the fast-takeoff debate, we formulate a conjunctive chain of jointly sufficient conditions for fast catastrophic supremacy. We operationalize these factors in a simulation of knowledge acquisition through hypothesis testing on NK fitness landscapes. Epistemic intelligence is represented by β, interpolating between random hypothesis selection and a greedy oracle. The simulations compare the effect of hypothesis-prioritization efficiency within scenarios that differ in initial knowledge, authorization barriers (control), and experimental latency. The simulations show that superior hypothesis prioritization reduces unproductive search without eliminating required validation time, while authorization barriers impose a greater proportional penalty on more efficient agents. The competence circuit identifies control and knowledge boundaries as primary levers for limiting agents' competence toward undesirable goals.

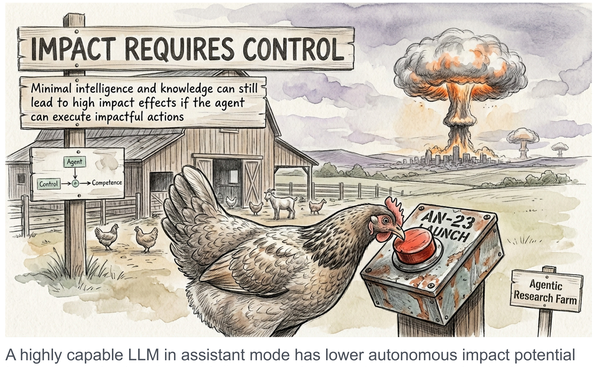
DEFINITIONS

- **Epistemic intelligence (I):** hypothesis-prioritization efficiency.
- **Knowledge (K):** knowledge available to the agent from the LLM and accessible resources (e.g., a knowledge base).
- **Control (C):** action surface available to the agent.

THE COMPETENCE CIRCUIT

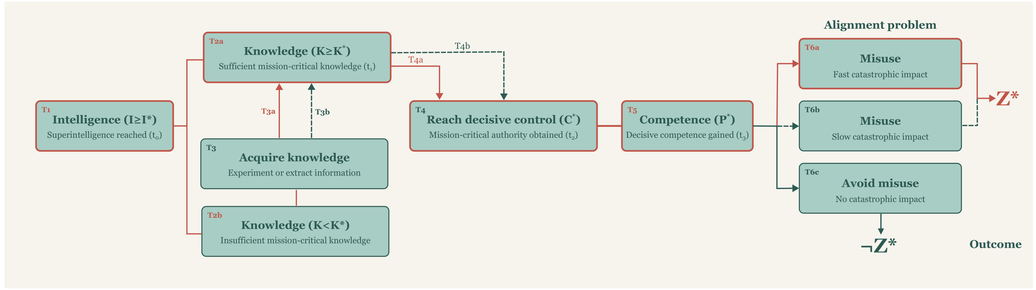


Graphical representation of the competence circuit: Competence = IKC + KC + C. Knowledge gates the impact of intelligence (see Archimedes argument at the end of poster), and control gates the contributions of both intelligence (IKC) and knowledge (KC) to competence while also contributing directly (C).



A highly capable LLM in assistant mode has lower autonomous impact potential than an agent with inferior intelligence but access to high-impact actuators.

SUFFICIENT CONDITIONS FOR FAST CATASTROPHIC SUPREMACY



The intelligence-mediated route (IKC) requires acquiring missing knowledge and obtaining mission-critical control. Existing know-how (KC) and direct control (C) provide additional routes. Red solid arrows denote fast transitions (minutes to weeks) and green dashed arrows denote slow transitions (months to years).

COMPUTATIONAL STRESS-TEST

An agent tests hypotheses on an NK fitness landscape. The simulation environments differ in initial knowledge cutoff, problem complexity (the proportion of unproductive alternatives and interactions among hypotheses), experimental verification latency, and authorization gates.

- **Intelligence β.** The agent selects the available hypothesis with the highest immediate gain per unit time cost with probability β and samples uniformly otherwise. β = 0 yields random search, while β = 1 yields a greedy oracle.
- **Experimental latency.** Every test takes time, including tests performed by the oracle. We represent latency by a fixed duration independent of β and compare fast digital and slow physical validation regimes.
- **Knowledge and control.** Initial knowledge determines which hypotheses are already eliminated. Authorization chokepoints add mandatory delays before granting access to downstream hypotheses.

KEY FINDINGS

- **Experimental latency limits the speed of knowledge acquisition.** Even the greedy oracle must perform the required tests to verify productive hypotheses. Eliminating unproductive search cannot eliminate that time.
- **Authorization barriers disproportionately delay superintelligent agents.** In the digital scenarios, authorization gates slow the superintelligent agent 4.9-fold (10.5 → 51.5) but the regular agent only 1.6-fold (90.3 → 146). Mandatory delays constitute a larger fraction of the efficient agent's otherwise shorter time budget, imposing a greater proportional penalty.
- **The fast-takeoff dilemma.** From Scenario 0 to 1 the intelligence multiplier rises from 1.8× to 7.5× because the regular agent wastes time on exploring unproductive hypotheses, while the superintelligent agent's exploration time barely changes. Large intelligence multipliers need large, noisy hypothesis spaces to confer a substantial advantage to superintelligence agents. In the physical domain, however, large hypothesis spaces coincide with slow experiments.

EFFECT OF INTELLIGENCE ON KNOWLEDGE-ACQUISITION EFFICIENCY

0 Connecting the dots
70% initial knowledge, no noise, weak hypothesis interactions, 1 time unit per test, no authorization gates.
Intelligence multiplier: 1.8× vs regular, 1.9× vs random.

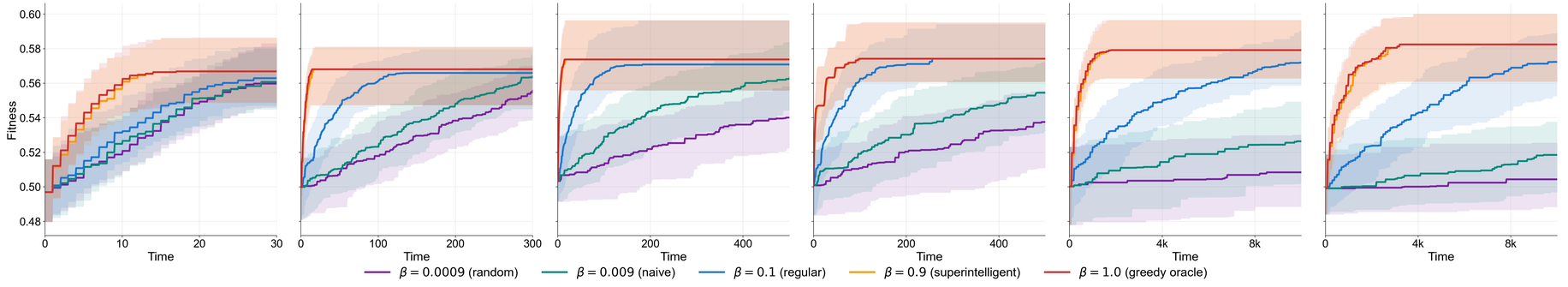
1 Disaster waiting to happen
Same conditions as 0, but with 3:1 noise.
Intelligence multiplier: 7.5× vs regular, 22.8× vs random.

2A Digital, weak safeguards
10% initial knowledge, 9:1 noise, moderate hypothesis interactions, 1 time unit per test, no authorization gates.
Intelligence multiplier: 8.6× vs regular, 77.2× vs random.

2B Digital, strong safeguards
Same conditions as 2A, plus 10 authorization gates. Gated tests take 20× the baseline test time.
Intelligence multiplier: 2.8× vs regular, 18.0× vs random.

3A Physical, weak safeguards
10% initial knowledge, 9:1 noise, stronger interactions, 100 time units per test, no authorization gates.
Intelligence multiplier: 8.9× vs regular, 82.4× vs random.

3B Physical, strong safeguards
Same conditions as 3A, plus 10 authorization gates. Gated tests take 5× the baseline test time.
Intelligence multiplier: 5.5× vs regular, 42.8× vs random.



Median fitness trajectories (lines) and interquartile ranges (bands) across 100 replications. The intelligence multiplier is the baseline agent's mean time to final fitness divided by the superintelligent agent's (β = 0.9) mean time to final fitness.

GOVERNANCE IMPLICATIONS

- **Knowledge boundaries.** Curate training and post-training data and compartmentalize access to information-retrieval tools and mission-critical sources.
- **Control boundaries.** Restrict executable actions through least privilege, isolation of systems and resources, and independent approval. These barriers address intelligence-mediated, procedural, and direct-control pathways, including when existing knowledge or fast digital testing makes discovery delays insufficient.
- **Social engineering.** Require separation of duties and limits on delegated authority so that a single manipulated operator cannot grant decisive control.
- **Experiments and extraction.** Sandbox agentic deployments and monitor and rate-limit system probing to slow digital experimentation and information extraction.
- **Physics in the loop.** Require multi-person human authorization or physical keys before mission-critical actions, placing an expensive control gate between digital decisions and execution.

POSITION

Control and knowledge boundaries are more decisive than intelligence limitation in constraining agents' competence toward undesirable goals.

ARCHIMEDES (287–212 BCE) ARGUMENT

Archimedes conceived the law of the lever and the concept of the center of gravity, derived the surface area and volume of the sphere, and laid foundations for integral calculus. Concluding that he was less intelligent than those who came later would contradict the evidence of his brilliance. He lacked later mathematical tools, such as algebraic notation, and the experimental findings on which modern physics rests. Likewise, superintelligence can improve hypothesis selection but remains constrained by available knowledge and the time required to obtain new empirical evidence.