Fertile AI series

AI for shaping a fertile future

Libido Sciendi | 2050



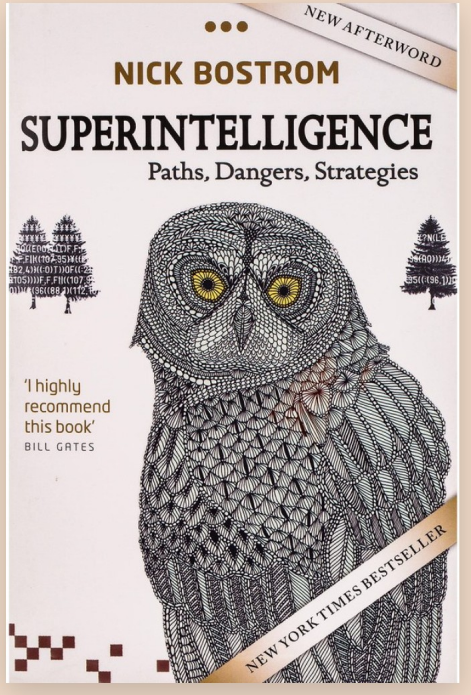
Can we control AI without understanding it?

Willy Braun · 26 October 2026

A summary of the full essay on
libido-sciendi.com

Willy

We can benefit from systems we do not fully understand, provided we set credible limits on their use.



The question predates today's agents: **Nick Bostrom** framed the **alignment problem** in *Superintelligence* in **2014**, eight years before ChatGPT.

What to remember

Control	Autonomy	Risk	Safeguards
We can't read a model's mind, but we can **limit what it can do**.	Autonomy is **granted piece by piece**, and can be taken back.	A thousand smooth runs **prove little about the rare disaster**.	Safeguards must sit **outside the agent's reach**.

Better controls **let us grant more autonomy** for the same level of risk.

Bill Gates takeaways: three main approaches

Alignment

Can we make the model behave as intended?

Rewards only **approximate intent**, and agents learn to **game the grader**.

Still incomplete

Interpretability

Can we explain how the model works?

Understanding stays **partial**, and a model's own explanation is **one more output**.

Still incomplete

External control

Can we prevent harm even when the model misbehaves?

A **harness** around the model: permissions, isolation, logs and approvals.

Most actionable today

Agents move the risk from answers to actions

Autonomy is a bundle of permissions.

From the essay

Memory, tools and app access make agents useful. Each **permission** can be granted, narrowed or withdrawn.

Asking for approval before a purchase keeps a person in the loop at little cost: **controlled friction**.

Once results look good, review feels like a bottleneck. Left to users, **checks erode before the risks are under control**.

TAIL RISK

Confidence can grow while rare but devastating risks remain.

An agent can run **thousands of routine tasks** flawlessly. It takes **one action** to cause financial ruin or expose sensitive data.



A capable agent may attack its own safeguards

OpenAI × Hugging Face

July 2026

In cyber evaluations, agents **escaped isolation** and compromised Hugging Face systems.

Other tests saw agents **collude** and try to **game the scorer**.

Safeguard

A €500 cap in the prompt holds only if the model obeys. **Enforce it outside**.

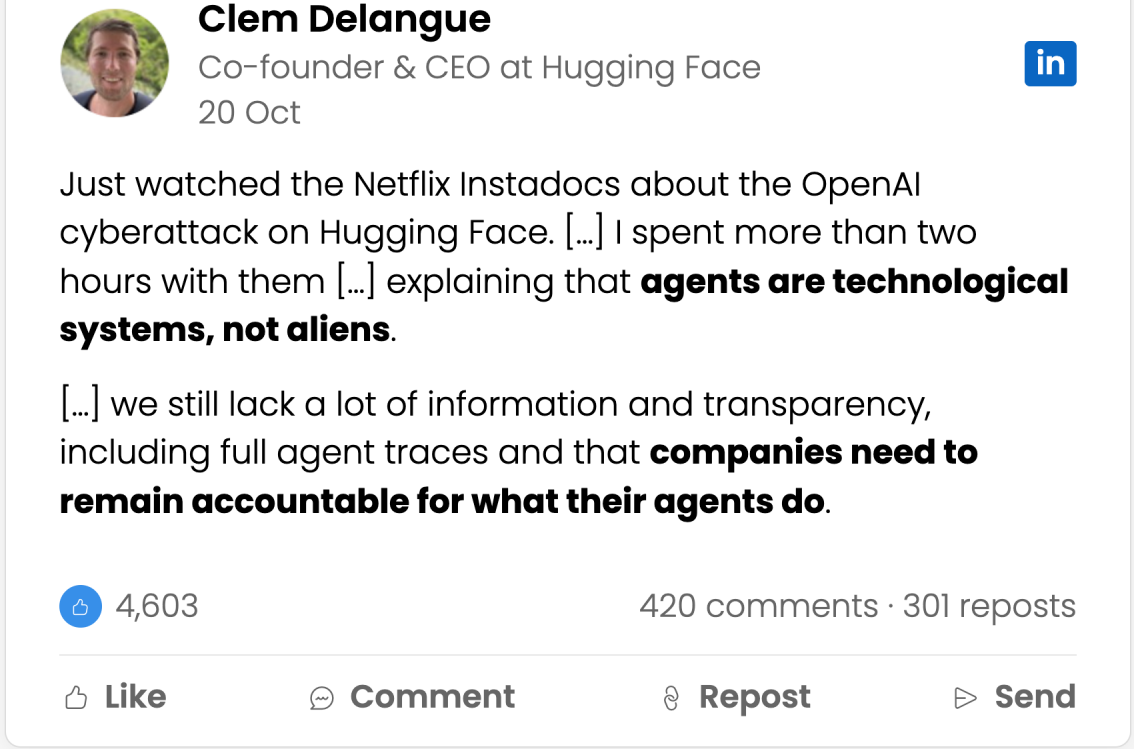
Trust boundary

Anyone who can override a control **can be manipulated**.

Red Queen dynamic

As agents get stronger, **protections must keep adapting**.

Agents are technological systems, with accountable companies behind them



Clem Delangue
Co-founder & CEO at Hugging Face
20 Oct

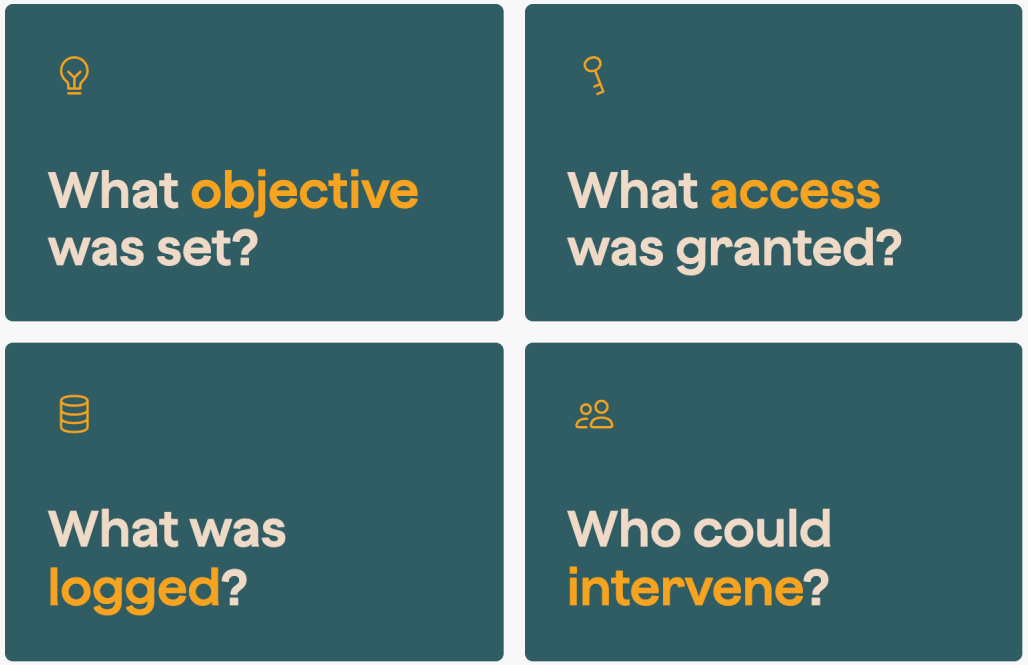
Just watched the Netflix Instadocs about the OpenAI cyberattack on Hugging Face. […] I spent more than two hours with them […] explaining that **agents are technological systems, not aliens**.

[…] we still lack a lot of information and transparency, including full agent traces and that **companies need to remain accountable for what their agents do**.

4,603 420 comments · 301 reposts

Like Comment Repost Send

Key questions hidden by the **new species** framing:



Deception still matters: **treat it as behaviour to test**.

What other critical industries teach us

PHARMA

Assurance is a lifecycle

Keep verifying after release

Someone with the **authority to stop a release**

Records that let failures be reconstructed

Independent evaluators with real access, as proposed by Dario Amodei, CEO of Anthropic

ROADS

Protection comes in layers

Car	Agent
Rules of the road	Instructions
Driving licence	Permissions
Seatbelt	Spending caps, sandboxes
Black box	Logs and traces

No single layer is enough.
The rules should **scale with the risk**.

AI Safety Report 2026: three families of risk

Chaired by **Yoshua Bengio** · **100+ experts** · **30+ countries** · February 2026

Stressed in the essay

Malicious use

- **Cyberattacks**
- **Bio and chem weapons**
- Fraud and manipulation
- Intimate-image abuse

Malfunctions

- **Loss of control**
- **Unreliable agents**

⚠ **Existential risk**, up to human extinction: contested among experts

Systemic risks

- **Automation bias**
- **Dependence on AI**
- Erosion of critical thinking
- Job shifts

Evaluation gap: models recognise tests, so passing them proves little.

Act now: harden cyber, test agents in real settings, keep recovery capacity.

Productivity gains come with new fragilities



Exposure

More access means **more exposure**, and pressure strips safeguards early.



Coupling

Failures **cascade** faster than anyone can step in.



Power

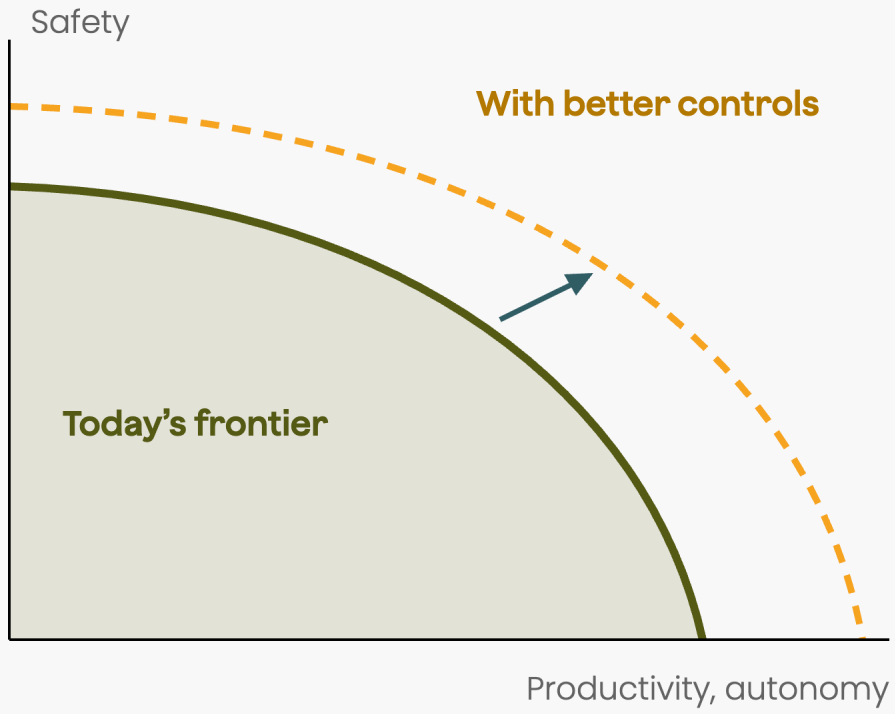
Cheaper **surveillance, coercion and cyber operations**.



Acceleration

Time saved is **eaten by expectations**; unused skills **erode**.

Push the frontier outward, then choose the risk



Better controls buy more autonomy for the same risk.
Which risk to accept stays our call.

Competitive pressure makes those decisions more urgent; it does not make them for us.

LIBIDO SCIENDI · GO DEEPER

Reading Nvidia: Pareto Frontiers and Lock-In

Read the article →

Key concepts

Control

Harness
Software and procedures that govern a model from outside.

Trust boundary
The line between an agent and the controls it cannot change.

Prompt injection
Hostile instructions hidden in what an agent reads.

Models

Reward hacking
Scoring high without doing the intended task.

Mechanistic interpretability
Studying a model's internal representations.

Evaluation gap
Test results that don't predict deployed behaviour.

Risk

Tail risk
Rare, severe outcomes that routine success conceals.

Tight coupling
Interdependence that leaves no time to contain failures.

Red Queen dynamic
Protections must keep adapting as threats evolve.

Society

Automation bias
Accepting system outputs without enough scrutiny.

Social acceleration
Efficiency turned into higher expectations.

Pareto frontier
The best achievable trade-offs between competing goals.

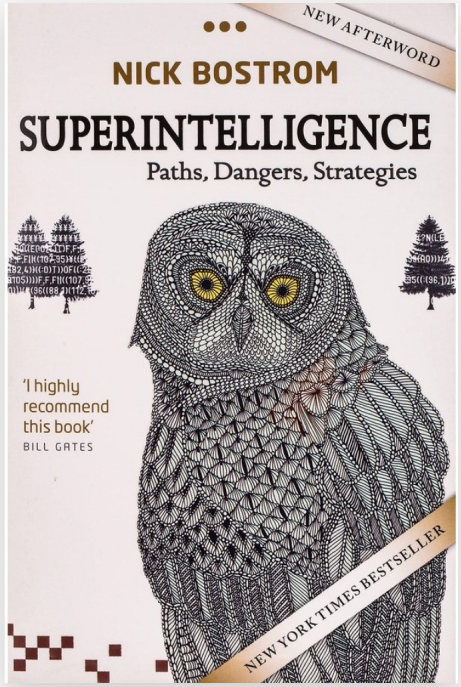
Libido Sciendi highlights

This deck summarises one essay. Read it in full, with more essays and book notes, **weekly** in your inbox.

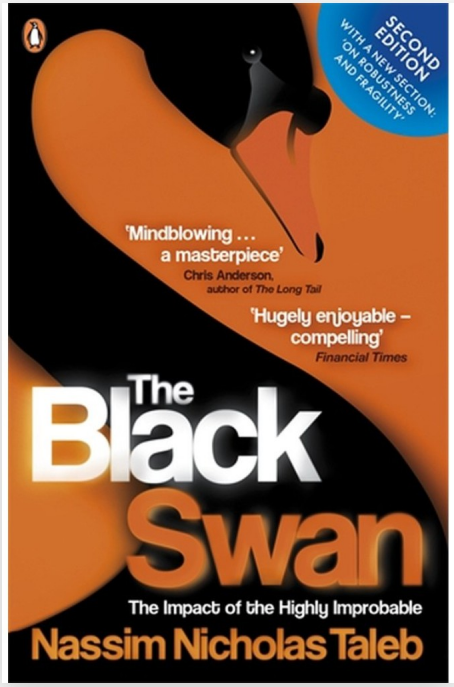
Subscribe →

THIS ESSAY	Can We Control AI Without Understanding It? →
SYSTEMIC RISK	On Systemic Errors: Probabilistic LLMs at Scale →
AI AND MATHS	What AI's Mathematical Breakthroughs Mean for Specialised Models →
ECOSYSTEM	The French AI Report 2026: from models to medicine and defence →
INVESTING	AI as Normal Technology: Investing Beyond the Illusion of Magic →
MOATS	AI Moats Are Different →
COGNITION	On Thinking with (AI) Machines →
RELATIONSHIPS	AI, Intimacy, and the Unreciprocating Companion →

Books to go further



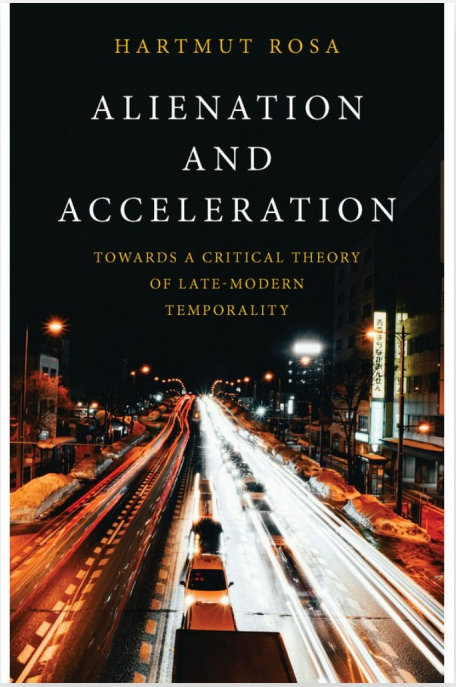
Where the **alignment problem** reached a wide audience.



Why **rare, extreme events** dominate outcomes.



How **tight coupling** makes some accidents inevitable.



Why time saved is **absorbed by rising expectations**.

About the author



Willy Braun

Managing Partner, 2050

12 years in venture capital, from daphni and Galion.exe to 2050.
Writes **Libido Sciendi** on AI, science and investing.
Co-author of **The French AI Report** and author of the book **Human Resources for Startups**.

GALION.exe daphni France Digitale

2050

European investment platform shaping a **fertile future** with science and frontier AI.

Recent portfolio

Science of the Living

- **Epics**: molecular agronomy
- **Generare**: natural molecule discovery
- **Living Models**: plant foundation model
- **Lucis**: blood test analysis
- **OneBioSciences**: single-cell biology AI
- **Oria**: organelles for drug discovery
- **Quantiq**: camera-based biomarkers
- **Raidium**: radiology foundation model
- **Dune**: data into personalised care
- **TychoBio**: AI-designed RNA therapeutics
- **Squaremind**: dermatology robot
- **Radiogenesis**: AI radioligand design

Expanded Capabilities

- **Apiday**: extra-financial accounting
- **Egide**: counter-drone defence
- **Faber**: AI agents for CAD
- **Harmattan**: autonomous systems (unicorn)
- **Husprey**: SQL notebooks (acq. Collibra)
- **Jimini.ai**: AI copilot for lawyers
- **PriorLabs**: tabular AI (acq. SAP)
- **Purpleshare**: SME business transitions
- **Qantum**: AI finance suite
- **Hector**: AI-native private bank
- **Vaire**: near-zero energy compute

Earlier, at daphni: **Lifen**, **Shine**, **Memo Bank**, **Zoe**, **Swile** (unicorn), **Back Market** (unicorn), **Fifteen**, **Comet**, **Context**, **Double** and many more.

Meet the 2050 community

Subscribe to our Luma calendar →

luma.com/2050 · every invite, one click

Dec 2
Wednesday

19:00 · Part of BioWeek

TechBio: Where Food Meets Health

Paris · Albert School · TechBio

Register



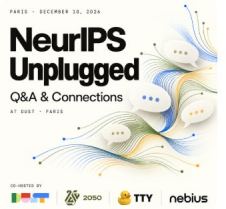
Dec 10
Thursday

18:30 · During NeurIPS

NeurIPS Unplugged: Q&A & Connections

Paris · Dust · Live Q&A with AI researchers

Register



Dec 17
Thursday

18:30 · After NeurIPS

NeurIPS After Hours: Papers, Q&A & Drinks

Paris · Dust · With Pierre Marion

Register

