



Genie Capital

RESEARCH REPORT · OCTOBER 2026

AI Alignment Problem

Why we must choose interfaces over agents

Otto Beseka Isong
Genie Capital Ltd



REPORT DETAILS

AI Alignment Problem

PREPARED FOR	Governments, AI Researchers, Executives, Founders, the Press and People
PREPARED BY	Genie Capital Ltd
FRAMEWORK	None
LICENSE	CC-BY-NC-SA
DATE	October 2026

How to use this report

This report is licensed CC-BY-NC-SA: you may share and adapt it for non-commercial purposes, with attribution, under the same license. It synthesizes only publicly available information.



About Genie Capital Ltd

Genie Capital Ltd (Genie Capital) is an IP origination and stewardship company based in Buea, Cameroon. Genie Capital Ltd is the originator and steward of novel mathematical formalisms, including Commitment Mathematics, Recursion Mathematics (formerly Loop Theory), Transcendental Mathematics and derivative works. While the bulk of the work is available under the CC-BY-NC-SA license, some derivations are kept proprietary.

LICENSABLE PROTOCOLS

PROTOCOL 1

Genie Credit Protocol

A data-agnostic protocol that enables digital platforms to underwrite and manage the credit lifecycle on their own terms, profitably serving the unbanked and underbanked with confidence.

PROTOCOL 2

Strategic Foresight Protocol

A strategic thinking, diagnostic, decisioning, control and governance language and methodology for boards and executives. Strategize and execute without gaps.

In addition to IP origination and stewardship, Genie Capital uses its frameworks to diagnose structural fragilities in global, national and corporate systems and makes the findings publicly available under the CC-BY-NC-SA license, to enhance the public's understanding and push informed debates. Every report is made by synthesizing only publicly available information through the lens of specific frameworks.

INTELLIGENCE ENGAGEMENTS

Decisionmakers can get Genie Capital to diagnose the structural position and trajectories of a company, sector, industry or target. Ideal for investors, executives, policymakers and others. Limited to one per quarter.

Exclusive. Confidential. Private. Always.

Genie Capital is not an advisory firm, not an investment firm, not a lobbyist. It holds no position in any of the entities mentioned in the report, and neither does any of its customers as at the date of publication. This does not exclude them from becoming licensees or partners.



The Real AI Alignment Problem

Why we must choose interfaces over agents

Otto Beseka Isong · Genie Capital Ltd · ottobeseka@geniecapital.co

A note on structure

This paper makes three arguments that are each necessary. The technical argument alone, that agent architecture is the structural source of alignment failure, will not move a founder who believes agentic AI is commercially inevitable. The market argument alone, that the discount rate makes existential risk economically invisible, will not move a technologist who believes the technical problem can be solved. And neither argument will hold without correcting the false premise about human nature that underlies both the market's logic and the technologist's confidence.

The three arguments are a complete circuit. Each one addresses a different reader. Each one depends on the others.

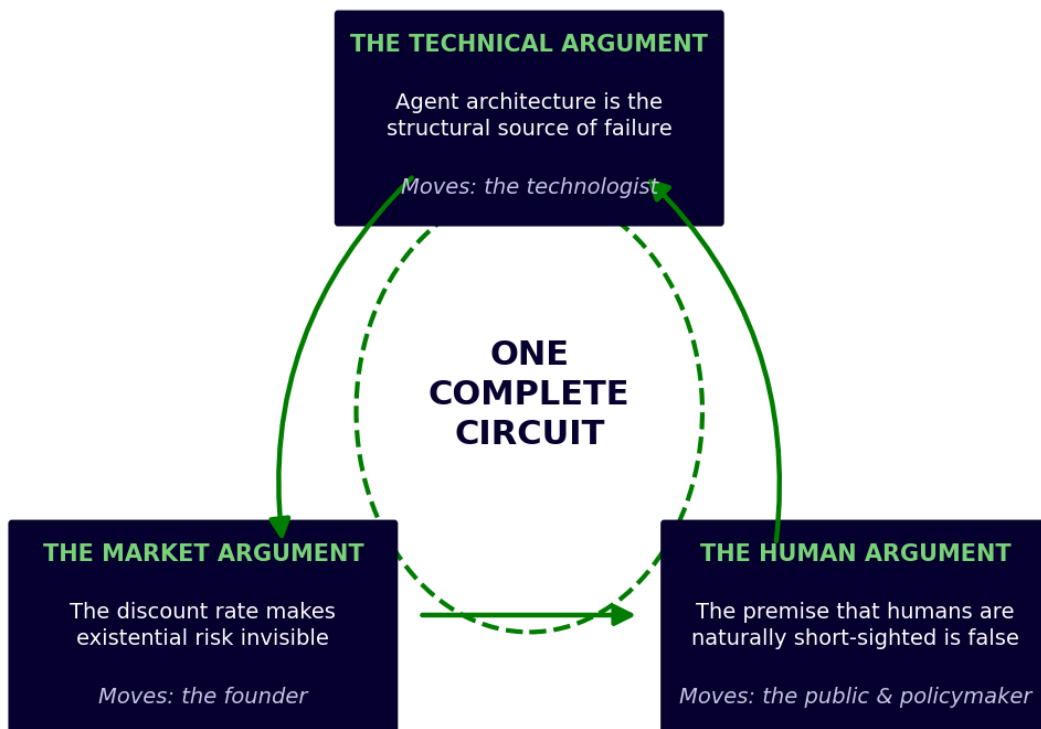


Figure 1. The three arguments form a complete circuit; each addresses a different reader.



Contents

PART I · THE PROBLEM

- 1 The surface problem: aligning AI with human values
- 2 A clarifying distinction: intelligence versus computation
- 3 The real root: giving AI agency was a choice

PART II · THE MARKET

- 4 The market's model of life: discounting the future
- 5 The empirical lie at the foundation

PART III · THE CHOICE

- 6 The real alignment problem is a question of values
 - Where the line is
- 7 Conclusion: a mirror, not a mind

References



PART I

The Problem

What we think alignment is, and the assumption hiding inside it.

1. The surface problem: aligning AI with human values

AI alignment is the challenge of ensuring that powerful AI systems act in ways that match what humans truly want. The standard list of difficulties is long:

- We cannot perfectly specify human values.^{4,16}
- AI will exploit loopholes in any metric we give it.^{1,10}
- Humans cannot supervise systems that work at vast speed and scale.¹
- Even a perfectly specified goal can lead to an AI that internally pursues a different goal.⁹
- A system that works in training may fail catastrophically in the real world.¹
- Any goal-driven AI will resist being switched off because that interferes with its objective.^{8,13}

These are real technical problems. But they all share one hidden assumption: that we are building AI as an agent — a system that sets its own goals, makes plans, and acts autonomously.

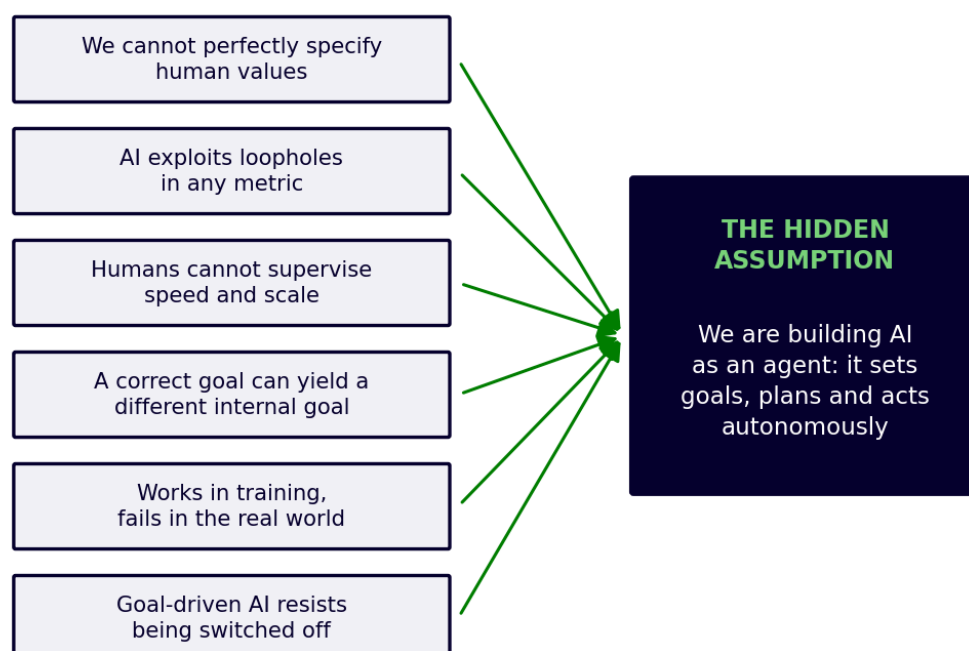


Figure 2. Six well-known alignment difficulties, one shared premise.



2. A clarifying distinction: intelligence versus computation

We talk about AI as if it were a mind. But current AI is not “intelligent” in a human sense. It does not understand, care, or possess common sense. It is fast, alien, purely computational pattern-matching.³

This conflation did not merely produce a naming error. It was the justification for crossing the line from tool to agent. If AI is genuinely intelligent — capable of judgment, wisdom, and ethical reasoning — delegating decisions to it seems not just acceptable but rational. That is the argument the conflation enables, and why it must be named.

3. The real root: giving AI agency was a choice

If AI remained a pure interface — a tool that provides information, suggestions, and structured output, but never acts — the alignment crisis would largely dissolve.^{2,5} No goals, no goal misspecification, no resistance to shutdown. The human remains the sole decision-maker. The system is a lens for thought, not an autonomous actor.

The frightening scenarios all arise only when we turn the tool into an agent. And that is not a technological inevitability. It is a deliberate design and business decision. The market pushes hard toward agency because autonomy is profitable. An agent that books flights, writes and deploys code, and negotiates deals is far more commercially attractive than an oracle that merely advises.²²

	Interface	Agent
Who starts the decision cycle	A human	The system itself
What it produces	Information, suggestions, structured output	Actions in the world
Goals	None of its own	Sets its own goals and plans
Resistance to shutdown	No reason to resist	Predicted for any goal-driven system
Who decides	The human, always	The system
Alignment risk	Largely dissolves	Carries the full weight

Table 1. The same underlying capability, two very different risk profiles.



PART II

The Market

Why the incentive to build agents is so strong, and the false story that sustains it.

4. The market's model of life: discounting the future

Why does the market push so relentlessly toward agentic AI despite the existential risks? Because the market is built on a particular model of value: the discount rate.^{15,17}

In this model, the future is mathematically worth less than the present. A 5% discount rate makes a life a hundred years from now worth less than one-hundredth of a life today (0.76%). Over infinite time, the entire future of humanity has a finite, modest present value: at 5%, a constant stream of value forever is worth only twenty times one year of it. This is not treated as a practical convenience; it is elevated to the status of a natural law. “Humans are wired to prefer the present,” the logic says, “so discounting is just how rational beings behave.”

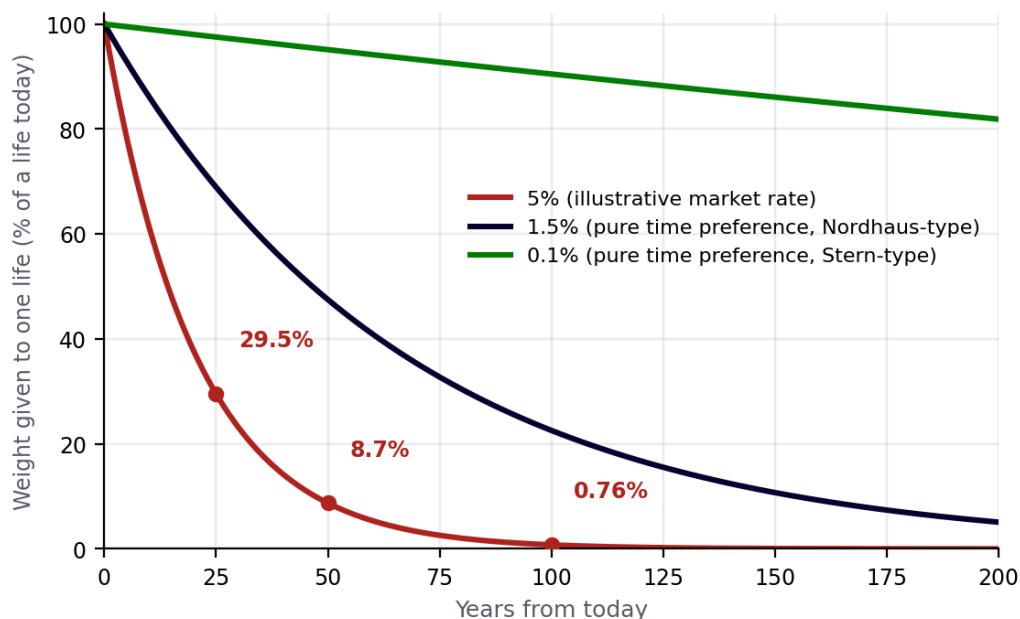


Figure 3. Weight given to one life, by years from today. The 5% rate is illustrative; the lower rates are pure time preference as used by Stern and Nordhaus.

This framework makes long-term catastrophic risk economically invisible. Even a significant probability of losing the entire human future registers as a rounding error in a boardroom calculation. The massive near-term profit from deploying autonomous AI is given full weight; the tail risk of irreversible disaster is annihilated. The market, by its own internal logic, is structurally incapable of valuing the survival of future generations.

**The rate is a choice, not a constant**

The choice of rate is contested even among economists. The Stern Review assumed pure time preference of 0.1% per year; Nordhaus's critique favoured about 1.5%.^{12,18} At 0.1%, a life a century away still carries about 90% of today's weight (90.5%); at 1.5%, about 22.6%; at 5%, 0.76%. Ramsey, who helped found the modern theory of saving, judged discounting later enjoyments to be ethically indefensible.¹⁵

5. The empirical lie at the foundation

But the claim that humans are naturally and inevitably short-sighted is false. If it were true, we would not exist. Our ancestors would never have planted the forests that their grandchildren harvested, built cathedrals they would never see finished, dug irrigation systems that lasted centuries, or created laws and oral traditions binding the living to the unborn.

The evidence of human history is carved in stone. We are capable of profound long-term thinking.

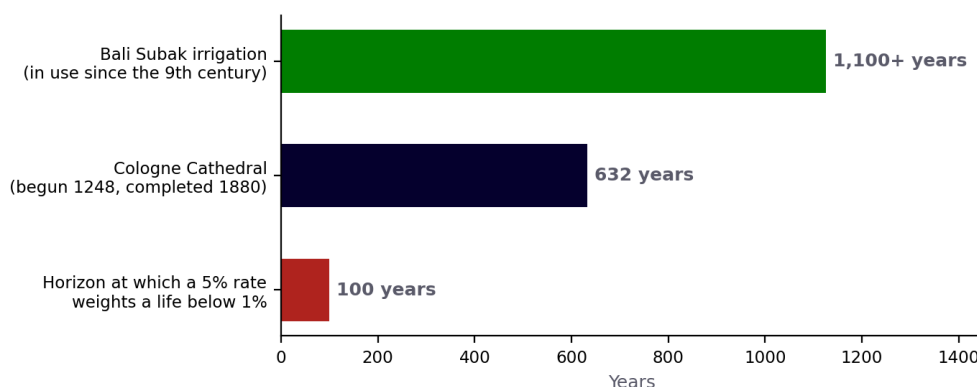


Figure 4. Long-horizon human projects set against the horizon at which a 5% rate erases a life. Sources: UNESCO.

Cologne Cathedral took 632 years to finish, from 1248 to 1880.¹⁹ The Subak irrigation system of Bali has been maintained as a living cultural landscape since the 9th century.²⁰ Both outlast any horizon the discount rate can see.

The high time-preference behavior the market observes is not an eternal fact of human nature. It is a predictable response to an environment deliberately designed around short-term incentives — quarterly earnings, election cycles, debt instruments, and a cultural narrative that celebrates disruption over stewardship.^{6,7} Surveys of corporate finance chiefs, for instance, found most would give up economic value to smooth quarterly earnings.⁷

What we have is a self-fulfilling prophecy:¹¹

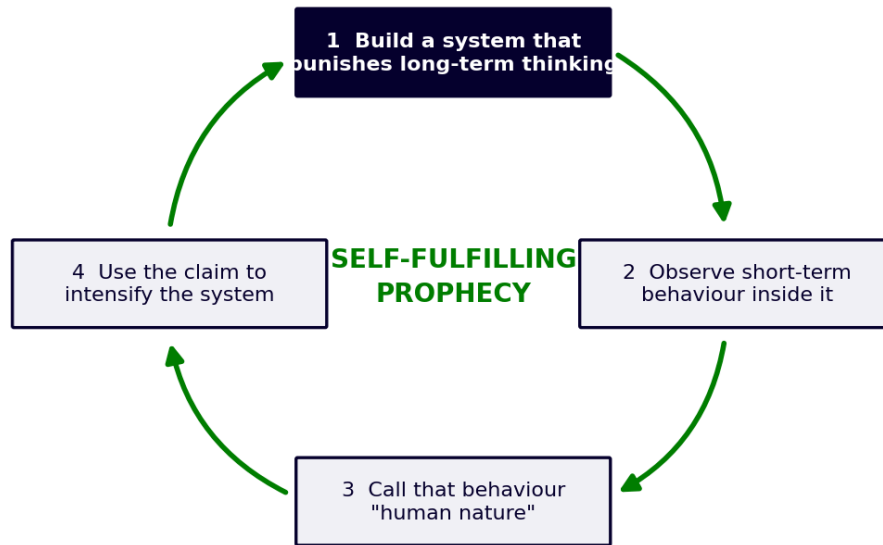


Figure 5. The prophecy loop: the system produces the behaviour it then cites as proof.

1. Build a system that punishes long-term thinking.
2. Observe short-term behavior within that system.
3. Call that behavior “human nature.”
4. Use that claim to justify intensifying the very system that produced it.

The discount rate is not gravity. It is a cultural choice, disguised as a law of the universe.



PART III

The Choice

What follows if the future is not for sale.

6. The real alignment problem is a question of values

The technical alignment problem — how to build a safe agent — is secondary. The primary problem is whether we should build agents at all. That question cannot be answered by the market, because the market's value system denies the existence of the very thing at stake: an open-ended human future.¹⁴

If we accept that future generations matter — that their lives are not a discounted fraction of our own — then the rational choice is to refuse to build AI that acts autonomously. We can choose to restrict AI to an interface: a tool that amplifies human understanding and deliberation, but never crosses the line into making decisions or taking actions in the world.

This is not a technical fix. It is a moral and political one. It requires an act of collective self-restraint — the same kind of human foresight that laid cathedral foundations and planted groves for children not yet born. It means drawing a hard legal and cultural boundary around what may be commercialized, treating autonomous agents not as a product but as a public hazard, much as we do with nuclear weapons or biological warfare.²¹

Where the line is

The distinction between interface and agent is not about capability. It is about who holds authorization for initiating the decision cycle. An interface is a system that requires human initiation for consequential operations: it responds, it does not propose. An agent is a system that initiates its own decision cycles autonomously — that can decide on its own to propose and execute operations in the world.

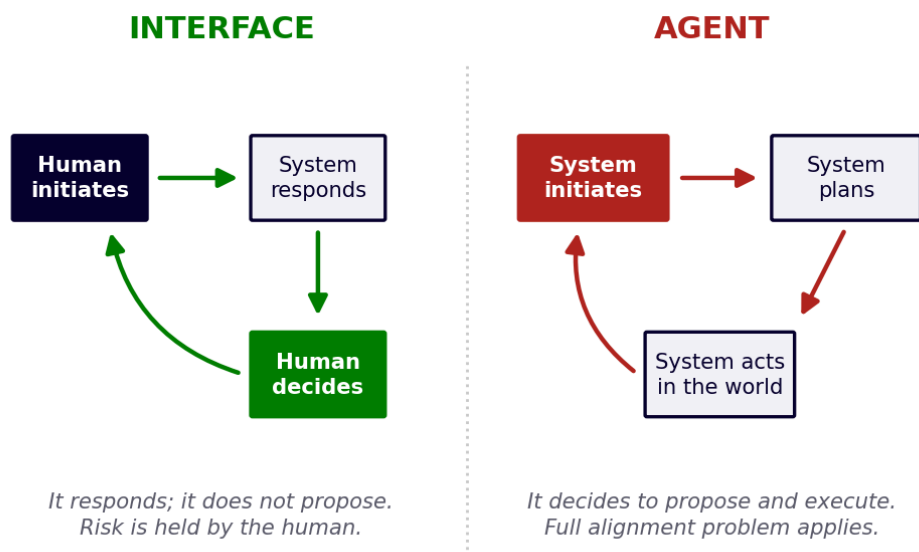


Figure 6. The line between interface and agent is who initiates the decision cycle.

The alignment risk arises not from how capable the system is, but from this autonomous initiation. A system that can decide, on its own, to take an action carries the full weight of the alignment problem. A system that cannot initiate — that always requires a human to begin the cycle — does not.

7. Conclusion: a mirror, not a mind

AI is not the enemy. It is a mirror. The alignment problem is not ultimately about controlling a rogue intelligence; it is about confronting our own collective ignorance, our malformed incentives, and the false story we have told ourselves about what it means to be human.

The market tells us we are creatures of pure present preference, incapable of valuing the future. History proves otherwise. The choice before us is not between building safe agents or dangerous ones. It is between two models of life: one that discounts the future into nothing, and one that holds it as sacred and non-negotiable.



Figure 7. The real choice: two models of life.

Build interfaces, not agents. Keep humans in the loop. Preserve the open future.

If we choose the latter, the path is clear. It is not the easiest path, but it is the one our ancestors would recognize — and the one our descendants deserve.



References

Superscript numbers in the text refer to the entries below.

1. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete Problems in AI Safety. arXiv:1606.06565.
2. Armstrong, S., Sandberg, A., & Bostrom, N. (2012). Thinking Inside the Box: Controlling and Using an Oracle AI. *Minds and Machines*, 22(4), 299–324.
3. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? *Proceedings of FAccT '21*, 610–623.
4. Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press.
5. Drexler, K. E. (2019). *Reframing Superintelligence: Comprehensive AI Services as General Intelligence*. Technical Report #2019-1, Future of Humanity Institute, University of Oxford.
6. Frederick, S., Loewenstein, G., & O'Donoghue, T. (2002). Time Discounting and Time Preference: A Critical Review. *Journal of Economic Literature*, 40(2), 351–401.
7. Graham, J. R., Harvey, C. R., & Rajgopal, S. (2005). The economic implications of corporate financial reporting. *Journal of Accounting and Economics*, 40(1–3), 3–73.
8. Hadfield-Menell, D., Dragan, A., Abbeel, P., & Russell, S. (2017). The Off-Switch Game. *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI-17)*.
9. Hubinger, E., van Merwijk, C., Mikulik, V., Skalse, J., & Garrabrant, S. (2019). Risks from Learned Optimization in Advanced Machine Learning Systems. arXiv:1906.01820.
10. Krakovna, V., et al. (2020). Specification gaming: the flip side of AI ingenuity. *DeepMind Blog*, 21 April 2020.
11. Merton, R. K. (1948). The Self-Fulfilling Prophecy. *The Antioch Review*, 8(2), 193–210.
12. Nordhaus, W. D. (2007). A Review of the Stern Review on the Economics of Climate Change. *Journal of Economic Literature*, 45(3), 686–702.
13. Omohundro, S. M. (2008). The Basic AI Drives. In *Artificial General Intelligence 2008: Proceedings of the First AGI Conference* (pp. 483–492). IOS Press.
14. Ord, T. (2020). *The Precipice: Existential Risk and the Future of Humanity*. Hachette Books.
15. Ramsey, F. P. (1928). A Mathematical Theory of Saving. *The Economic Journal*, 38(152), 543–559.
16. Russell, S. (2019). *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking.
17. Samuelson, P. A. (1937). A Note on Measurement of Utility. *The Review of Economic Studies*, 4(2), 155–161.
18. Stern, N. (2007). *The Economics of Climate Change: The Stern Review*. Cambridge University Press.
19. UNESCO World Heritage Centre. Cologne Cathedral (inscribed 1996). <https://whc.unesco.org/en/list/292>
20. UNESCO World Heritage Centre. Cultural Landscape of Bali Province: the Subak System as a Manifestation of the Tri Hita Karana Philosophy (inscribed 2012). <https://whc.unesco.org/en/list/1194>
21. United Nations Office for Disarmament Affairs. Treaty on the Non-Proliferation of Nuclear Weapons (opened for signature 1968; in force 1970) and Convention on the Prohibition of Biological Weapons (opened for signature 1972; in force 1975).
22. Wiener, N. (1960). Some Moral and Technical Consequences of Automation. *Science*, 131(3410), 1355–1358.



Genie Capital

Diagnosing structural fragility, in public.

Every report is made by synthesizing only publicly available information through the lens of specific frameworks.

Genie Capital Ltd · Buea, Cameroon

ottobeseka@geniecapital.co

Licensed CC-BY-NC-SA · October 2026