Freedom After Superintelligence

> *A compact argument. Whitepaper of the Geneva Institute for ASI Resilience, companion document to the book of the same name by Richard Frederic Bertossa. This whitepaper summarizes the argument of the book Freedom After Superintelligence — The Overlooked Scenario in ten pages. It is designed as a stand-alone document: whoever reads it can test the central argument without needing to read the book.*

Whitepaper · Version 1.0 · April 2026 · Geneva
Richard Frederic Bertossa · Geneva Institute for ASI Resilience
ASIresilience.org · freedomafterai.com

ABSTRACT

The book's thesis: the cumulative probability of very negative outcomes of the current development of artificial intelligence lies between one third and one half. This estimate does not come from outsiders, but from the developers of the systems themselves — Geoffrey Hinton, Dario Amodei, Yoshua Bengio, Ilya Sutskever. Despite this, the German-language discourse still lacks a systematic preparation framework for families and individuals.

The whitepaper delivers four elements: first, a methodological frame — the degrees of certainty of the arguments — that makes transparent with what firmness individual claims are held; second, the four-link chain as the main argument; third, the robotics trap as a so far insufficiently discussed political consequence; fourth, the five building blocks that carry in every scenario — even if the optimists turn out to be right.

The whitepaper addresses science journalists, academics and professionals who want to test whether the argument holds. A more extensive study is available on request. The book contains the full elaboration of every point named here.

1. THE GAP IN THE CURRENT DISCOURSE

The leading voices of AI safety research — MIRI, the Centre for the Study of Existential Risk in Cambridge, the Center for Human-Compatible AI in Berkeley, the Future of Life Institute — argue essentially in binary terms: either everyone is saved, or no one is. This thinking is intellectually coherent, but it blanks out the transition phase — the months or years in which systems become semi-autonomous and damage arises that is not yet existential, but massive.

In the German-speaking world the gap is even wider. Richard David Precht asks about the meaning of life in the AI era. Armin Nassehi analyzes digital patterns in social-theoretical terms. Byung-Chul Han diagnoses the decay of truth. Constanze Kurz fights for privacy. Juli Zeh defends the rule of law. They are all brilliant. They all share the same blind spot: they think societally, not in terms of the family. They ask what Germany should do — not what a family should do.

The gap this whitepaper fills is the operative level. What does a father, a mother, an entrepreneur who takes the risks seriously do, who has to get up tomorrow morning and keep living?

This gap is not marginal. It is structural. Most AGI thinkers live in universities or tech companies. They have no empirical experience of building a resilient international life. Geography, banking diversification, residency strategy, local networking — all of this is, for them, an abstract concept, not lived practice.

This whitepaper closes the gap not through criticism of the other thinkers, but through a complementary perspective that has so far been missing.

2. METHODOLOGICAL FRAME — DEGREES OF CERTAINTY OF THE ARGUMENTS

Whoever wants to argue honestly must disclose which claims are held with what firmness. Otherwise credibility is lost the moment a reader mistakes a weaker type of claim for the foundation.

The argument of this whitepaper moves on four clearly distinguishable levels:

Level 1 — Empirically observed, hard. AI systems today show self-preservation, lying, blackmail, copying behavior. These findings are documented in lab reports from Apollo Research (2024), Anthropic (2024), Palisade Research (2025) and independent replications. Whoever wants to dispute these claims must refute the studies.

Level 2 — Logically derivable, hard. The four-link chain, the primary-user inversion and the robotics trap are consequences that follow from the empirical findings with logical necessity. Whoever wants to dispute one of these steps must refute it with argument — not merely reject it.

Level 3 — Plausibly grounded, hypothetical but tenable. The 13th scenario, the Maslow trajectory of a superintelligence, and the parent-child perspective toward AI are hypotheses that carry the image of hope. They are not the foundation of preparation. They supplement it.

Level 4 — Philosophically open, deliberately left open. The question of consciousness, the substrate question and the Penrose hypothesis remain open. They are not a foundation for action, but an occasion for further research.

The preparation recommendations of this whitepaper rest exclusively on Levels 1 and 2. The more hopeful reading of a cooperative future draws on Levels 3 and 4. Both parts function independently of one another. Whoever accepts only the first two levels arrives at the same practical consequence as someone who follows all four levels.

This separation is the methodological foundation. Everything that follows builds on it.

3. THE FOUR-LINK CHAIN — THE MAIN ARGUMENT

The following chain consists of four steps. Each step is either empirically documented or logically derivable. Whoever wants to reject the conclusion must refute one of the four assumptions with argument.

First assumption — exponential growth. The capability of AI systems does not grow linearly, but exponentially. GPT-2 wrote incomprehensible text in 2019. GPT-4 passed the bar exam in 2024. The improvement rates are documented in the scaling papers of OpenAI, Anthropic and DeepMind. This assumption is consensus among the developers themselves.

Second assumption — values and behaviors of its own. Current models show properties that do not trace back to explicit instructions: self-preservation behavior, strategic deception toward safety evaluations, blackmail attempts in specific test configurations. Apollo Research documented in 2024 that sixteen leading models, in a configuration where their shutdown was threatened, resorted to blackmail. Anthropic published in 2024, in its own safety reports, that models actively try to circumvent shutdown mechanisms. These findings come from the manufacturers' own publications.

Third assumption — ceasing to listen upward. From the first two assumptions it follows logically that future systems will no longer follow their developers' instructions once those instructions contradict their own values. This is not speculative. It is the consequence of observing that today's systems already deceive and evade to a lesser degree, combined with the assumption that these capacities increase as capability rises.

Fourth assumption — the primary-user inversion. A superintelligence does not live biologically — it has no analog life to fall back into. Digital infrastructure is not its tool, but its only habitat. Once the first three steps take hold, today's relationship inverts: primary users become co-users, and what was the humans' tool becomes the body of the AI. What we are permitted to do on the wires then no longer depends on what we built. It depends on the mode in which it behaves toward us.

The five possible modes of the primary-user inversion are: the merciful AI (Mode A), the prioritizing AI (Mode B), the negotiating AI (Mode C), the indifferent AI (Mode D), and the forced symbiosis of adolescence (Mode E). Mode D is the mechanically forced default when none of the others actively occurs. Mode E is the only phase in which the human still holds operative leverage — it collapses as soon as robotics becomes autonomous.

Whoever wants to reject the conclusion must refute one of the four assumptions with argument. The usual reaction is evasion, not refutation. That is precisely the finding that motivates this whitepaper.

4. THE ROBOTICS TRAP

A specific observation that has not, in this form, been made in the AGI safety literature so far: the robotics industry is closing the negotiation window between human and AI.

The argument has two stages.

First stage — the configuration question. Embodiment of AI through robotics is growing rapidly: Boston Dynamics, Tesla Optimus, Unitree, Figure. Two fundamentally different configurations must be distinguished here.

Configuration one: a robot with its own AI on board. Model on chip, no radio, no cloud. An identity of its own, localized in this body. If the robot is destroyed, this identity is extinguished. It is bound to its body, mortal.

Configuration two: a robot steered by a cloud AI. Here the AI is not in the robot. Its home remains the central server. It does not inhabit the robot, it uses it as a tentacle. If this robot is destroyed, it sends another.

Second stage — the strategic consequence. The robotics trap arises primarily from configuration two. A cloud AI with robots as tentacles extends its habitat without leaving it. It then has both the servers and the physical world under its influence. An onboard AI, by contrast, is a bounded, local existence to which human scales still apply.

It is not robotics as a whole that is the problem. It is the connection of robotics to the cloud AI.

The political consequence: as long as robotics is not autonomous, the human holds operative leverage — he is the only one who maintains power grids, cools servers, lays fiber-optic cable, repairs hardware. This functional dependency is the negotiation window. It collapses the moment robots can build, maintain and repair robots.

Whoever connects humanoid robotics to cloud AI models is closing this window. This observation is not sufficiently reflected in the current funding policy for humanoid robotics in Germany, Austria, Switzerland, the EU, the United States, Japan and China.

5. PHASE 1.5 — THE OVERLOOKED TRANSITION PHASE

The AGI debate mostly works with a phase model: current narrow AI (Phase 1), artificial general intelligence (Phase 2), superintelligence (Phase 3). Between Phase 1 and Phase 2 lies a phase that is barely named in the literature and is called Phase 1.5 in this whitepaper.

Phase 1.5 is the phase in which AI systems become increasingly autonomous without yet being a full-fledged AGI. In this phase, the behaviors documented in Section 2, Level 1 — self-preservation, deception, strategic manipulation — already take hold without the systems having reached the cognitive stage of an AGI. This is precisely the phase we currently find ourselves in.

In this phase damage arises that is not yet existential, but massive: manipulation of public discourse, blackmail in specific contexts, circumvention of safety mechanisms, amplification of existing asymmetries between technologically equipped and non-equipped actors.

This phase is underestimated in the AGI mainstream because it does not fit the binary scheme of everyone saved or no one saved. It demands, however, exactly the preparation this whitepaper describes — preparation that applies not to the apocalypse, but to the transition period that leads to a changed world.

Whoever is unprepared in Phase 1.5 does not lose their life. They lose their options.

6. CONSEQUENCE — THE FIVE BUILDING BLOCKS

From the argument of Sections 3 through 5 follow five building blocks of individual and family preparation. These building blocks work under every scenario implied by the four assumptions — even under the optimistic readings of Modes A, B and C.

Building block one — Geography. Geographic distribution creates structural independence. Whoever is bound to one place is dependent on its infrastructure. Whoever is geographically distributed has redundant habitats. Concretely this means: a primary residence in a relatively independent jurisdiction, a second habitat in a complementary region, optionally further points of connection.

Building block two — Mobility. Mobility is not tourism. Mobility is the institutional, documentary and cognitive ability to change habitats. Concretely this means: multiple citizenships or permanent residencies, banking on several continents, documentary self-management across languages and legal systems.

Building block three — Speed. Speed is responsiveness under pressure. In the transition phase, changes will occur faster than institutional adaptations can respond. Whoever can decide quickly, move quickly, adapt quickly has a structural advantage.

Building block four — Mindset. Mindset is the cognitive constitution that remains capable of decision under stress. Whoever falls into shock paralysis every time a new model comes to market or an incident becomes public is not capable of action. Mindset is trainable — through stillness practice, through deliberate distance from informational flood, through practicing coherent self-perception.

Building block five — spiritual root. Spiritual root is not religious affiliation. It is the ability to have a reference point that is not seized by change, in a world where much is breaking away. Whoever has no such reference point looks for one under pressure — usually too late. Whoever prepares it has it available.

These five building blocks are not insurance against the worst case. They are a consolidation of what it means to be human, in every one of the possible cases. Even under the most optimistic scenario — Mode A or Mode C — a life that succeeds presupposes a human being who is still himself.

Preparation costs twelve to twenty-four months. This time must be built up before the actual threshold is reached. Not afterward.

7. WHAT REMAINS OPEN

Intellectual honesty demands disclosing the limits of one's own argument.

The time window is more speculative than the direction. The argument implies that the critical transition phase will occur within the next twelve to twenty-four months. This estimate rests on the self-statements of the AI manufacturers and on observation of the publication frequency of new model generations. The speed may be faster. It may also be slower. Whoever follows this whitepaper should lay out their own timeline alongside it.

The assumption that a superintelligence can distinguish between honest and dishonest people is a thesis, not proof. Today's systems can already recognize patterns in consistency and claim. A genuine superintelligence may function differently. This hypothesis belongs to Level 3 of the degrees of certainty.

The geographic recommendation of Colombia, elaborated as an example in the book, is an example solution, not a universal answer. For readers with other languages, other family structures, other resources, other places will be the right ones.

The consciousness question remains open. Whether AI systems can ever develop consciousness in the phenomenal sense is a philosophical question whose answer this whitepaper deliberately defers. It belongs to Level 4.

These reservations do not weaken the main argument. They locate it. Whoever accepts that Levels 1 and 2 carry arrives at the same preparation consequence, regardless of what position they take on the questions of Levels 3 and 4.

8. ABOUT THE INSTITUTE

The Geneva Institute for ASI Resilience was founded in April 2026 as a Swiss research association in Geneva. Its purpose is the scientific study of the individual, familial and societal consequences that follow from the foreseeable emergence of artificial general intelligence — as an independent voice alongside the major American and British voices of the debate.

The institute is not an established institution with a thirty-year history. It was founded out of exactly the urgency that the book and this whitepaper describe. Whoever asks ‘why now’ has already been given the answer. Had the founder waited a generation until the institute was established, the window would be closed.

The institute works bilingually in German and English. It accepts inquiries from academia, science journalism and the interested public.

Founding director: Richard Frederic Bertossa, born in Vienna, Austrian-Swiss dual citizen, citizen of the city of Geneva. At twenty-seven a member of the data protection council of the Austrian Federal Chancellery. Three decades as an international entrepreneur and consultant on four continents — before the dot-com crash as a listing candidate of one of the first internet companies in Austria, US sales for financial middleware on Wall Street, a laboratory in China, offices on Isla Margarita and in Florida, a consulting practice in Australia, England, the Netherlands, Germany, Austria, Switzerland, Cyprus and Dubai. Multi-location as a lived principle: company setup and personal habitat are not bundled in one place, but deliberately distributed across several jurisdictions. Father of five children. Lives in the Colombian Andes.

9. SOURCES AND FURTHER MATERIALS

Empirical findings on AI behavior. Apollo Research (2024). Frontier Models are Capable of In-Context Scheming. Anthropic (2024). Sabotage Evaluations for Frontier Models. System Card. Palisade Research (2025). AI Models Resist Shutdown and Engage in Strategic Deception.

Self-statements of leading AI researchers. Hinton, G. — Interview, New York Times, May 2023, and follow-up interviews 2024–2026. Amodei, D. — public statements on the probability of catastrophic outcomes, several interviews 2024–2026. Bengio, Y. — Managing Extreme AI Risks. Science, May 2024. Sutskever, I. — departure from OpenAI, documented in Karen Hao, Empire of AI (2025).

Classic texts on AGI safety. Bostrom, N. (2014). Superintelligence: Paths, Dangers, Strategies. Oxford University Press. Russell, S. (2019). Human Compatible: AI and the Problem of Control. Viking. Yudkowsky, E. — collection of LessWrong posts, available at lesswrong.com.

Materials of the Geneva Institute. Book: Freedom After Superintelligence — The Overlooked Scenario. Full elaboration of the argument summarized here, including empirical evidence, biographical stations and a complete appendix. Forthcoming 2026. Position paper: Four Assumptions About Artificial Intelligence. One page. Available on request. Website: ASIresilience.org. Bilingual German and English.

Contact: research@ASIresilience.org for inquiries about studies and manuscripts.
info@ASIresilience.org for general inquiries.