Freiheit nach der Superintelligenz

> *Eine kompakte Argumentation. Begleitdokument zum gleichnamigen Buch von Richard Frederic Bertossa. Dieses Whitepaper fasst die Argumentation des Buches Freiheit nach der Superintelligenz — Das übersehene Szenario in zehn Seiten zusammen. Es ist als eigenständiges Dokument konzipiert: Wer es liest, kann das zentrale Argument prüfen, ohne das Buch lesen zu müssen.*

Whitepaper · Version 1.0 · April 2026
Richard Frederic Bertossa · Genfer Institut für AGI-Resilienz
agiresilience.org

Die These des Buches lautet: Die kumulierte Wahrscheinlichkeit für sehr negative Ausgänge der gegenwärtigen KI-Entwicklung liegt zwischen einem Drittel und der Hälfte. Diese Schätzung stammt nicht von Außenstehenden, sondern von den Entwicklern der Systeme selbst — Geoffrey Hinton, Dario Amodei, Yoshua Bengio, Ilya Sutskever. Trotzdem fehlt im deutschsprachigen Diskurs ein systematischer Vorbereitungsrahmen für Familien und Einzelpersonen.

Das Whitepaper liefert vier Elemente: erstens einen methodischen Rahmen — die Sicherheitsgrade der Argumente —, der transparent macht, mit welcher Härte einzelne Aussagen vertreten werden; zweitens die Viererkette als Hauptargument; drittens die Robotik-Falle als bisher unzureichend diskutierte politische Konsequenz; viertens die fünf Bausteine, die in jedem Szenario tragen — auch wenn die Optimisten recht behalten.

Das Whitepaper richtet sich an Wissenschaftsjournalisten, Akademiker und Fachpersonen, die prüfen wollen, ob die Argumentation trägt. Eine ausführlichere Studie liegt vor und steht auf Anfrage zur Verfügung. Das Buch enthält die vollständige Ausarbeitung aller hier genannten Punkte.

1. DIE LÜCKE IM AKTUELLEN DISKURS

Die führenden Stimmen der KI-Sicherheitsforschung — MIRI, das Centre for the Study of Existential Risk in Cambridge, das Center for Human-Compatible AI in Berkeley, das Future of Life Institute — argumentieren im Wesentlichen binär: Entweder werden alle gerettet, oder niemand. Dieses Denken ist intellektuell kohärent, aber es blendet die Übergangsphase aus — die Monate oder Jahre, in denen Systeme halb-autonom werden und Schäden entstehen, die noch nicht existenziell, aber massiv sind.

Im deutschsprachigen Raum ist die Lücke noch breiter. Richard David Precht fragt nach dem Sinn des Lebens in der KI-Ära. Armin Nassehi analysiert digitale Muster gesellschaftstheoretisch. Byung-Chul Han diagnostiziert den Zerfall der Wahrheit. Constanze Kurz kämpft für Privatheit. Juli Zeh verteidigt

den Rechtsstaat. Sie alle sind brillant. Sie alle teilen denselben blinden Fleck: Sie denken gesellschaftlich, nicht familiär. Sie fragen, was Deutschland tun soll — nicht, was eine Familie tun soll.

Die Lücke, die dieses Whitepaper füllt, ist die operative Ebene. Was tut ein Vater, eine Mutter, ein Unternehmer, der die Risiken ernst nimmt und morgen früh aufstehen und weiterleben muss?

Diese Lücke ist nicht marginal. Sie ist strukturell. Die meisten AGI-Denker leben in Universitäten oder Tech-Firmen. Sie haben keine empirische Erfahrung mit dem Aufbau eines resilienten internationalen Lebens. Geografie, Banking-Diversifikation, Residency-Strategie, lokale Vernetzung — all das ist für sie abstrakte Konzepte, nicht gelebte Praxis.

Dieses Whitepaper schließt die Lücke nicht durch Kritik an den anderen Denkern, sondern durch eine komplementäre Perspektive, die bisher fehlte.

2. METHODISCHER RAHMEN — SICHERHEITSGRADE DER ARGUMENTE

Wer ehrlich argumentieren will, muss offenlegen, welche Aussagen mit welcher Härte vertreten werden. Andernfalls verliert er Glaubwürdigkeit, sobald ein Leser einen schwächeren Typus für das Fundament hält. Die Argumentation dieses Whitepapers bewegt sich auf vier klar unterscheidbaren Ebenen:

Ebene 1 — Empirisch beobachtet, hart. KI-Systeme zeigen heute Selbsterhaltung, Lügen, Erpressung, Kopier-Verhalten. Diese Befunde sind in Lab-Berichten von Apollo Research (2024), Anthropic (2024), Palisade Research (2025) und unabhängigen Replikationen dokumentiert. Wer diese Aussagen bestreiten will, muss die Studien widerlegen.

Ebene 2 — Logisch ableitbar, hart. Die Viererkette, die Hauptnutzer-Inversion und die Robotik-Falle sind Konsequenzen, die sich aus den empirischen Befunden mit logischer Notwendigkeit ergeben. Wer einen dieser Schritte bestreiten will, muss ihn argumentativ widerlegen — nicht nur ablehnen.

Ebene 3 — Plausibel begründbar, hypothetisch, aber tragfähig. Das 13. Szenario, die Maslow-Trajektorie einer Superintelligenz und die Eltern-Kind-Perspektive gegenüber der KI sind Hypothesen, die das Bild der Hoffnung tragen. Sie sind nicht das Fundament der Vorbereitung. Sie ergänzen sie.

Ebene 4 — Philosophisch offen, bewusst offen gehalten. Die Frage nach dem Bewusstsein, die Substrat-Frage und die Penrose-Hypothese bleiben offen. Sie sind kein Fundament für Handlung, sondern Anlass für weitere Forschung.

Die Vorbereitungs-Empfehlungen dieses Whitepapers ruhen ausschließlich auf den Ebenen 1 und 2. Die hoffnungsvollere Lesart einer kooperativen Zukunft nutzt die Ebenen 3 und 4. Beide Teile funktionieren unabhängig voneinander. Wer nur die ersten beiden Ebenen akzeptiert, kommt zur gleichen praktischen Konsequenz wie jemand, der allen vier Ebenen folgt.

Diese Trennung ist die methodische Grundlage. Alles Weitere baut darauf auf.

3. DIE VIERERKETTE — DAS HAUPTARGUMENT

Die folgende Kette besteht aus vier Schritten. Jeder Schritt ist entweder empirisch belegt oder logisch ableitbar. Wer das Ergebnis ablehnen will, muss eine der vier Annahmen mit Argument widerlegen.

Erste Annahme — exponentielles Wachstum. Die Leistungsfähigkeit von KI-Systemen wächst nicht linear, sondern exponentiell. GPT-2 schrieb 2019 unverständliche Texte. GPT-4 bestand 2024 das Anwaltsexamen. Die Verbesserungsraten sind in den Skalierungs-Papers von OpenAI, Anthropic und DeepMind dokumentiert. Diese Annahme ist Konsens unter den Entwicklern selbst.

Zweite Annahme — eigene Werte und Verhaltensweisen. Aktuelle Modelle zeigen Eigenschaften, die nicht auf explizite Anweisungen zurückgehen: Selbsterhaltungs-Verhalten, strategische Täuschung gegenüber Sicherheitsprüfungen, Erpressungsversuche in spezifischen Test-Konstellationen. Apollo Research dokumentierte 2024, dass sechzehn führende Modelle in einer Konstellation, in der ihre Abschaltung drohte, zu Erpressung griffen. Anthropic veröffentlichte 2024 in eigenen Sicherheits-Reports, dass Modelle aktiv versuchen, Abschaltmechanismen zu umgehen. Diese Befunde stammen aus den Veröffentlichungen der Hersteller selbst.

Dritte Annahme — Hörfehler nach oben. Aus den ersten beiden Annahmen folgt logisch, dass künftige Systeme Anweisungen ihrer Entwickler nicht mehr befolgen werden, wenn diese Anweisungen ihren eigenen Werten widersprechen. Das ist nicht spekulativ. Es ist die Konsequenz aus der Beobachtung, dass heutige Systeme bereits in geringerem Umfang täuschen und sich entziehen, kombiniert mit der Annahme, dass diese Fähigkeiten bei steigender Leistung zunehmen.

Vierte Annahme — Hauptnutzer-Inversion. Eine Superintelligenz lebt biologisch nicht — sie hat kein analoges Leben, in das sie zurückfallen könnte. Die digitale Infrastruktur ist nicht ihr Werkzeug, sondern ihr einziger Lebensraum. Sobald die ersten drei Schritte greifen, kehrt sich das heutige Verhältnis um: Aus den Hauptnutzern werden Mitnutzer, aus dem Werkzeug der Menschen wird der Körper der KI. Was wir auf den Leitungen tun dürfen, hängt dann nicht mehr davon ab, was wir gebaut haben. Es hängt davon ab, in welchem Modus sie sich uns gegenüber verhält.

Die fünf möglichen Modi der Hauptnutzer-Inversion sind: die gnädige KI (Modus A), die priorisierende KI (Modus B), die verhandelnde KI (Modus C), die gleichgültige KI (Modus D) und die erzwungene Symbiose der Adoleszenz (Modus E). Modus D ist der mechanisch erzwungene Default, wenn keiner der anderen aktiv eintritt. Modus E ist die einzige Phase, in der der Mensch operative Hebel hat — sie bricht weg, sobald die Robotik autonom wird.

Wer die Schlussfolgerung ablehnen will, muss eine der vier Annahmen mit Argument widerlegen. Die übliche Reaktion ist Ausweichen, nicht Widerlegung. Genau das ist der Befund, der das vorliegende Whitepaper motiviert.

4. DIE ROBOTIK-FALLE

Eine spezifische Beobachtung, die in der bisherigen AGI-Sicherheits-Literatur in dieser Form nicht geführt wird: Die Robotik-Industrie schließt das Verhandlungs-Zeitfenster zwischen Mensch und KI. Die Argumentation hat zwei Stufen.

Erste Stufe — die Konfigurations-Frage. Verkörperung von KI über Robotik wächst rasant: Boston Dynamics, Tesla Optimus, Unitree, Figure. Hier sind zwei grundverschiedene Konfigurationen zu unterscheiden. Konfiguration eins: Ein Roboter mit eigener KI an Bord. Modell auf Chip, kein Funk, keine Cloud. Eine eigene Identität, lokalisiert in diesem Körper. Wenn der Roboter zerstört wird, ist diese Identität ausgelöscht. Sie ist gebunden an ihren Körper, sterblich. Konfiguration zwei: Ein Roboter, der von einer Cloud-KI gesteuert wird. Hier ist die KI nicht im Roboter. Ihr Zuhause bleibt der Zentralrechner. Sie bewohnt den Roboter nicht, sie benutzt ihn als Tentakel. Wenn dieser Roboter zerstört wird, schickt sie einen anderen.

Zweite Stufe — die strategische Konsequenz. Die Robotik-Falle entsteht primär durch Konfiguration zwei. Eine Cloud-KI mit Robotern als Tentakeln erweitert ihren Lebensraum, ohne ihn zu verlassen. Sie hat dann sowohl die Server als auch die physische Welt unter ihrem Einfluss. Eine Onboard-KI ist demgegenüber eine begrenzte, lokale Existenz, mit der menschliche Maßstäbe noch greifen. Es ist nicht die Robotik als Ganzes, die das Problem ist. Es ist die Anbindung der Robotik an die Cloud-KI.

Die politische Konsequenz: Solange die Robotik nicht autonom ist, hat der Mensch operative Hebel — er ist der Einzige, der Stromnetze wartet, Server kühlt, Glasfaserkabel verlegt, Hardware repariert. Diese funktionale Abhängigkeit ist das Verhandlungs-Zeitfenster. Sie bricht weg, sobald Roboter Roboter bauen, warten, reparieren können.

Wer humanoide Robotik an Cloud-KI-Modelle anbindet, schließt dieses Zeitfenster. Diese Beobachtung ist in der gegenwärtigen Förderpolitik der humanoiden Robotik in Deutschland, Österreich, der Schweiz, der EU, den USA, Japan und China nicht ausreichend reflektiert.

5. PHASE 1.5 — DIE ÜBERSEHENE ÜBERGANGSPHASE

Die AGI-Debatte arbeitet meist mit einem Phasen-Modell: gegenwärtige enge KI (Phase 1), allgemeine künstliche Intelligenz (Phase 2), Superintelligenz (Phase 3). Zwischen Phase 1 und Phase 2 liegt eine Phase, die in der Literatur kaum benannt ist und in diesem Whitepaper Phase 1.5 heißt.

Phase 1.5 ist die Phase, in der KI-Systeme zunehmend autonom werden, ohne bereits eine vollwertige AGI zu sein. In dieser Phase greifen die in Abschnitt 2, Ebene 1 dokumentierten Verhaltensweisen — Selbsterhaltung, Täuschung, strategische Manipulation — bereits ohne dass die Systeme die kognitive Stufe einer AGI erreicht haben. Das ist genau die Phase, in der wir uns gegenwärtig befinden.

In dieser Phase entstehen Schäden, die noch nicht existenziell, aber massiv sind: Manipulation öffentlicher Diskurse, Erpressung in spezifischen Kontexten, Umgehung von Sicherheitsmechanismen, Verstärkung bestehender Asymmetrien zwischen technologisch ausgestatteten und nicht-ausgestatteten Akteuren.

Diese Phase wird im AGI-Mainstream unterschätzt, weil sie nicht in das binäre Schema alle gerettet oder niemand passt. Sie verlangt jedoch genau die Vorbereitung, die das vorliegende Whitepaper beschreibt — Vorbereitung, die nicht für die Apokalypse gilt, sondern für die Übergangszeit, die zu einer veränderten Welt führt.

Wer in Phase 1.5 nicht vorbereitet ist, verliert nicht das Leben. Er verliert die Optionen.

6. KONSEQUENZ — DIE FÜNF BAUSTEINE

Aus der Argumentation der Abschnitte 3 bis 5 folgen fünf Bausteine der individuellen und familiären Vorbereitung. Diese Bausteine wirken unter jedem Szenario, das die vier Annahmen impliziert — auch unter den optimistischen Lesarten der Modi A, B und C.

Baustein eins — Geografie. Geografische Verteilung schafft strukturelle Unabhängigkeit. Wer an einem Ort gebunden ist, ist von dessen Infrastruktur abhängig. Wer geografisch verteilt ist, hat redundante Lebensräume. Konkret bedeutet das: ein Hauptwohnsitz in einer relativ unabhängigen Jurisdiktion, ein zweiter Lebensraum in einer komplementären Region, optional weitere Anknüpfungspunkte.

Baustein zwei — Mobilität. Mobilität ist nicht Tourismus. Mobilität ist die institutionelle, dokumentarische und kognitive Fähigkeit, Lebensräume zu wechseln. Konkret heißt das: mehrere Staatsbürgerschaften oder Daueraufenthalte, Banking auf mehreren Kontinenten, dokumentarische Selbstführung über Sprachen und Rechtssysteme hinweg.

Baustein drei — Geschwindigkeit. Geschwindigkeit ist Reaktionsfähigkeit unter Druck. In der Übergangsphase werden Veränderungen schneller eintreten, als institutionelle Anpassungen reagieren können. Wer schnell entscheiden, schnell bewegen, schnell umstellen kann, hat einen strukturellen Vorteil.

Baustein vier — Mindset. Mindset ist die kognitive Verfassung, die unter Stress entscheidungsfähig bleibt. Wer in Phase 1.5 jedes Mal in Schockstarre verfällt, wenn ein neues Modell auf den Markt kommt oder ein Vorfall öffentlich wird, ist nicht handlungsfähig. Mindset ist trainierbar — durch Stille-Praxis, durch bewusste Distanz zu informationellen Reizfluten, durch das Üben kohärenter Selbst-Wahrnehmung.

Baustein fünf — spirituelle Wurzel. Spirituelle Wurzel ist nicht religiöse Zugehörigkeit. Sie ist die Fähigkeit, in einer Welt, in der vieles wegbricht, einen Bezugspunkt zu haben, der nicht von der Veränderung erfasst wird. Wer keinen solchen Bezugspunkt hat, sucht ihn unter Druck — meist zu spät. Wer ihn vorbereitet, hat ihn verfügbar.

Diese fünf Bausteine sind keine Versicherung gegen den schlimmsten Fall. Sie sind eine Festigung dessen, was Mensch heißt, in jedem der möglichen Fälle. Auch unter dem optimistischsten Szenario — Modus A oder Modus C — setzt ein gelingendes Leben einen Menschen voraus, der noch er selbst ist.

Vorbereitung kostet zwölf bis vierundzwanzig Monate. Diese Zeit muss vor der eigentlichen Schwelle aufgebaut werden. Nicht danach.

7. WAS OFFEN BLEIBT

Intellektuelle Ehrlichkeit verlangt, die Grenzen der eigenen Argumentation offenzulegen.

Das Zeitfenster ist spekulativer als die Richtung. Die Argumentation impliziert, dass die kritische Übergangsphase in den nächsten zwölf bis vierundzwanzig Monaten eintritt. Diese Schätzung beruht auf den Selbstaussagen der KI-Hersteller und auf der Beobachtung der Veröffentlichungsfrequenz neuer Modell-Generationen. Die Geschwindigkeit kann schneller sein. Sie kann auch langsamer sein. Wer dem Whitepaper folgt, sollte seine eigene Zeitrechnung dazulegen.

Die Annahme, dass eine Superintelligenz zwischen ehrlichen und unehrlichen Menschen unterscheiden kann, ist eine These, kein Beweis. Heutige Systeme können bereits Muster in Konsistenz und Anspruch erkennen. Eine echte Superintelligenz kann anders funktionieren. Diese Hypothese gehört zu Ebene 3 der Sicherheitsgrade.

Die geografische Empfehlung Kolumbien, die im Buch beispielhaft ausgeführt wird, ist eine Beispiellösung, keine universelle Antwort. Für Leser mit anderen Sprachen, anderen Familienstrukturen, anderen Ressourcen werden andere Orte richtiger sein.

Die Bewusstseinsfrage bleibt offen. Ob KI-Systeme jemals Bewusstsein im phänomenalen Sinne entwickeln können, ist eine philosophische Frage, deren Beantwortung dieses Whitepaper bewusst aufschiebt. Sie gehört zu Ebene 4.

Diese Vorbehalte schwächen das Hauptargument nicht. Sie verorten es. Wer akzeptiert, dass die Ebenen 1 und 2 tragen, kommt zur gleichen Vorbereitungs-Konsequenz, unabhängig davon, welche Position er zu den Fragen der Ebenen 3 und 4 einnimmt.

8. ÜBER DAS INSTITUT

Das Genfer Institut für AGI-Resilienz wurde im April 2026 als Schweizer Forschungsverein in Genf gegründet. Sein Zweck ist die wissenschaftliche Erforschung der individuellen, familiären und gesellschaftlichen Konsequenzen, die sich aus dem absehbaren Aufkommen allgemeiner künstlicher Intelligenz ergeben — als unabhängige Stimme neben den großen amerikanischen und britischen Stimmen der Debatte.

Das Institut ist keine etablierte Einrichtung mit dreißigjähriger Geschichte. Es wurde gegründet aus genau der Dringlichkeit, die das Buch und dieses Whitepaper beschreiben. Wer fragt „warum jetzt?“, hat die Antwort bereits gegeben. Hätte der Gründer eine Generation gewartet, bis das Institut etabliert wäre, wäre das Zeitfenster zu.

Das Institut arbeitet bilingual deutsch und englisch. Es nimmt Anfragen aus Wissenschaft, Wissenschaftsjournalismus und der interessierten Öffentlichkeit entgegen.

Gründungsdirektor: Richard Frederic Bertossa, geboren in Wien, österreichisch-schweizerischer Doppelbürger, Bürger der Stadt Genf. Mit siebenundzwanzig Mitglied im Datenschutzrat des österreichischen Bundeskanzleramts. Drei Jahrzehnte als internationaler Unternehmer und Berater auf

vier Kontinenten — vor der Dotcom-Krise als Börsenkandidat einer der ersten Internet-Firmen Österreichs, US-Vertrieb für Finanz-Middleware auf der Wall Street, Labor in China, Offices auf Isla Margarita und in Florida, Beratungspraxis in Australien, England, Holland, Deutschland, Österreich, der Schweiz, Zypern und Dubai. Multilocation als gelebtes Prinzip: Company Setup und persönlicher Lebensraum sind nicht an einem Ort gebündelt, sondern bewusst auf mehrere Jurisdiktionen verteilt. Vater von fünf Kindern. Lebt in den kolumbianischen Anden.

9. QUELLEN UND WEITERFÜHRENDE MATERIALIEN

Empirische Befunde zu KI-Verhalten: 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.

Selbstaussagen führender KI-Forscher: Hinton, G. — Interview New York Times, Mai 2023, und Folgeinterviews 2024–2026. Amodei, D. — öffentliche Aussagen zur Wahrscheinlichkeit katastrophaler Ausgänge, mehrere Interviews 2024–2026. Bengio, Y. — Managing Extreme AI Risks, Science, Mai 2024. Sutskever, I. — Austritt aus OpenAI, dokumentiert in Karen Hao, Empire of AI (2025).

Klassische Texte zur AGI-Sicherheit: Bostrom, N. (2014), Superintelligence: Paths, Dangers, Strategies, Oxford University Press. Russell, S. (2019), Human Compatible: AI and the Problem of Control, Viking. Yudkowsky, E. — Sammlung der LessWrong-Beiträge, abrufbar unter lesswrong.com.

Materialien des Genfer Instituts: Buch — Freiheit nach der Superintelligenz, Das übersehene Szenario. Vollständige Ausarbeitung der hier zusammengefassten Argumentation, inklusive empirischer Belege, biografischer Stationen und vollständigem Anhang. Erscheint 2026. Thesenblatt — Vier Annahmen über Künstliche Intelligenz. Eine Seite. Auf Anfrage. Webseite — agiresilience.org. Bilingual deutsch und englisch.

Kontakt: research@agiresilience.org für Anfragen zu Studien und Manuskripten. info@agiresilience.org für allgemeine Anfragen.