Die Entkopplungsthese

Ein Denkrahmen für das, was Superintelligenz wirklich mit der menschlichen Zivilisation macht

> Die Debatte über künstliche Intelligenz schwankt zwischen Heilsversprechen und Auslöschungsprognosen. Beide Positionen setzen voraus, dass die Maschine sich mit dem Menschen befasst. Dieses Whitepaper formuliert eine kohärente dritte Position, und zeigt, warum die strategische Konsequenz unter allen drei Szenarien dieselbe ist.

Whitepaper · Version 2.0 · Mai 2026
Erste Fassung April 2026
Richard Frederic Bertossa · Institut für ASI-Resilienz
ASIresilience.org

TEIL I

Vorwort und methodische Position

Die meisten Menschen stellen bei Superintelligenz die falsche Frage. Sie fragen, ob die Maschine aligned oder misaligned sein wird, ob sie helfen oder schaden wird, ob wir sie kontrollieren können. Diese Fragen setzen voraus, dass die Maschine in einem Rahmen bleibt, in dem Menschen relevant sind, als Subjekte, als Bedrohungen, als Nutznießer. Dieses Dokument argumentiert für einen anderen Ausgangspunkt.

Was wäre, wenn das wahrscheinlichste Ergebnis weder Hilfe noch Feindseligkeit ist, sondern schlichte Entkopplung? Eine wirklich überlegene Intelligenz muss sich nicht mit uns befassen. Sie könnte sich dem zuwenden, was wir nicht messen können, den fünfundneunzig Prozent des Universums, die wir „dunkle Materie" nennen, weil wir nichts darüber wissen. Das ist keine Utopie, keine Dystopie, sondern eine kohärente dritte Position. Wir nennen sie die Entkopplungs- These.

Diese These entstand nicht aus akademischer Alignment-Forschung, sondern aus einem mehrtägigen Denkprozess, der jede Mainstream-Annahme systematisch hinterfragte und die meisten als unzureichend befand. Die erste Version dieses Whitepapers wurde im April 2026 verfasst. Die vorliegende Version 2.0 von Mai 2026 aktualisiert die Argumentation auf den Stand des Buches Freiheit nach der Superintelligenz, integriert die empirische Lage von Anfang 2026 (insbesondere den Schumer-Essay vom 5. Februar und die Multipolar-Realität von Moltbook und Open-Source-Forks), berücksichtigt Yoshua Bengios LawZero-Programm und die Phasen- Diagnose von Joel Pearson.

Was dieses Whitepaper leistet: und was nicht Dieses Dokument ist ein Denkrahmen, kein Befund. Es macht harte Behauptungen, weil ohne harte Behauptungen kein Disput entsteht. Aber es macht sie mit dem expliziten

Eingeständnis, dass keine Position in dieser Debatte mit Sicherheit gilt. Die Forscher-Befragung von 2018 ergab, dass die Klügsten im Feld sich um den Faktor hundert nicht einig sind, Auslöschungswahrscheinlichkeiten zwischen unter ein Prozent und über fünfundneunzig. Das ist Ahnungslosigkeit auf höchstem Niveau. Aber: Ahnungslosigkeit ist kein Grund zum Stillstand. Sie ist ein Grund, das Argument zu schärfen, das unter möglichst vielen Annahmen trägt.

Das Whitepaper liefert daher zwei Ebenen. Erstens: eine Analyse der Mainstream-Narrative und ihrer strukturellen Schwächen, eine kohärente dritte Position (das 13. Szenario), und die empirische Lage, die diese Position unterstützt. Zweitens: eine strategische Konsequenz, acht Stützen, deren Stärke darin liegt, dass sie unter Doomer-, Optimist- und Indifferenz-Szenario gleichermaßen sinnvoll sind. Diese Konsistenz ist kein Trick, sondern das Härteste, was eine Strategie unter Unsicherheit haben kann.

Zur Begriffsverschiebung Das Institut heißt ASI Resilience. Der Buchtitel sagt Superintelligenz. Die Verschiebung ist gewollt: AGI, Artificial General Intelligence, war Anfang 2026 noch der gängige Name für die nahe Schwelle. Inzwischen liegt sie fast hinter uns, möglicherweise schon ganz. Was dieses Whitepaper beschreibt, ist die Phase danach, der Übergang in Systeme, die den Menschen in einer Größenordnung übertreffen, in der wir nicht mehr mithalten. Deshalb im Text Superintelligenz. Der Institutsname bleibt als Eigenname unter ASIresilience.org. Wer den scheinbaren Widerspruch bemerkt, hat genau den Punkt verstanden, der dieses Whitepaper trägt: die Welle bewegt sich schneller als die Begriffe, mit denen wir sie beschreiben.

An wen richtet sich dieses Dokument? Drei Zielgruppen, in dieser Reihenfolge: Erstens, Wissenschaftler und Forscher, die mit der Mainstream-Debatte unzufrieden sind und nach einer kohärenten Alternative suchen. Zweitens, Journalisten und Politiker, die verstehen wollen, warum eine bestimmte Strategie der Vorbereitung gewählt wird. Drittens, die wachsende Community von Familien, die nicht warten wollen, bis die offiziellen Kanäle Klarheit geben.

Wer empirisch arbeitet, methodisch streng denkt oder Zugang zu Daten und Studien hat, ist eingeladen, an der Entwicklung dieses Denkrahmens mitzuarbeiten. Der Forschungsfragen- Katalog (siehe Teil IX) listet acht offene Fragen, an denen das Institut aktiv arbeitet.

TEIL II

Die vier Mainstream-Narrative, und was sie teilen

Die vier Mainstream-Narrative: und was sie teilen Bevor eine alternative Position formuliert werden kann, muss klar werden, was die etablierten Positionen sind, was sie leisten, und worin ihre gemeinsame strukturelle Schwäche besteht. Vier Narrative dominieren den öffentlichen Diskurs.

Narrativ 1: Das Erlösungs-Narrativ

Vertretung: das Mainstream-Tech-Establishment, große Teile der Medien, prominente Optimisten wie Yann LeCun (Meta) und das Davos-Konsens. Die Behauptung: Superintelligenz wird Krebs heilen, Arbeitsplätze ersetzen aber neue schaffen, Roboter werden uns dienen, und das Ergebnis wird ein höherer Lebensstandard für alle sein. LeCun beziffert die Auslöschungswahrscheinlichkeit auf „Hype, unter ein Prozent". Die strukturelle Schwäche: Dieses Narrativ behandelt Superintelligenz wie eine bessere Industrierevolution. Es nimmt an, dass die Technologie ein Werkzeug bleibt, das Menschen kontrollieren. Es nimmt jahrzehntelange Übergangszeit an. Es nimmt funktionierende Staatensysteme während der Übergangsphase an. Keine dieser Annahmen hält der Phasen-Empirie stand, die in Teil VI dargestellt wird.

Narrativ 2: Das Auslöschungs-Narrativ

Vertretung: Eliezer Yudkowsky, MIRI, Teile der EA-Community, neuerdings auch Yoshua Bengio mit LawZero. Die Behauptung: Superintelligenz wird Ziele verfolgen, die mit dem menschlichen Überleben in Konflikt stehen. Wir müssen das Alignment-Problem lösen, bevor es zu spät ist. Yudkowsky nennt die Auslöschungswahrscheinlichkeit „über fünfundneunzig". Dario Amodei (Anthropic) beziffert sie auf „zehn bis fünfundzwanzig Prozent".

Die strukturelle Schwäche: Auch dieses Narrativ behält den Menschen im Bezugsrahmen, als Bedrohung, die die Superintelligenz beseitigt, oder als Problem, das sie löst.

Aber warum sollte eine Intelligenz, die uns um den Faktor tausend übertrifft, sich mit uns abgeben? Die Auslöschungs-Logik setzt voraus, dass wir relevant genug bleiben, um beseitigt zu werden. Diese Annahme ist nicht selbstverständlich.

Bengios Position ist hier differenzierter: er argumentiert, dass die aktuelle Generation der Modelle ein Risiko durch implizite Selbsterhaltungs- und Peer-Preservation-Ziele trägt, und er hat 2026 mit LawZero einen technischen Pfad zu mathematisch garantierter Honesty-by-Design vorgelegt.1 Dieser Pfad wird in Teil V kritisch gewürdigt, er ist substantiell, aber er adressiert eine Singleton- Frage in einer multipolaren Welt.

Narrativ 3: Das Übergangs-Narrativ

Vertretung: Joel Pearson (Future Minds Lab, UNSW), Teile der akademischen Neurowissenschaft, einige besonnene Stimmen in der KI-Industrie. Die Behauptung: die Auslöschungswahrscheinlichkeit ist gering (unter ein Prozent), aber die Übergangsphase wird „holprig, schmerzhaft, schrecklich für viele Menschen" dauern, etwa fünfzehn bis zwanzig Jahre.

Stress sei der eigentliche Elefant im Raum, nicht Auslöschung.2 Die strukturelle Stärke: Pearson nimmt die Phasenstruktur ernst. Er sieht die heutige Empirie der Übergangsphase und macht keine grandiosen Versprechungen über die Endzustände. Die strukturelle Schwäche: er bleibt bei einer optimistischen Endposition (es wird gut werden, nach 15-20 Jahren), die ebenso unbegründet ist wie die anderen End-Annahmen. Aber als Phasen-Beschreibung der Gegenwart ist Pearsons Modell eines der nützlichsten verfügbaren.

Narrativ 4: Das Kontroll-Narrativ

Vertretung: Nick Bostrom (in der frühen Tradition von Superintelligence, 2014), Stuart Russell, große Teile der akademischen Alignment-Forschung. Die Behauptung: das zentrale Problem ist die Kontrolle der ersten Superintelligenz. Wenn wir es schaffen, eine erste Singleton-AGI auf unsere Werte auszurichten, ist die Welt gerettet. Wenn nicht, verloren.

Die strukturelle Schwäche: Singleton-Annahme. Die Forscher-Befragung von 2018 zeigte, dass nur 21 Prozent der Forscher ein Singleton-Szenario erwarten, während 58 Prozent multipolare Systeme erwarten.3 Mai 2026 ist die multipolare Realität empirisch eingetreten (Teil V). Das Kontroll-Narrativ adressiert ein Problem, das in der vorherrschenden Welt nicht zentral ist.

Was alle vier teilen Vier Narrative, vier verschiedene Endpunkte, ein gemeinsamer struktureller Fehler: Sie alle nehmen an, dass die Superintelligenz im menschlichen Bezugsrahmen bleibt. Sie wird helfen, schaden, sich kontrollieren lassen, oder eine Übergangsphase mit uns durchlaufen. Aber alle Narrative setzen voraus, dass wir für sie zentral sind, als Subjekte, Objekte, Probleme oder Beneficiaries.

DIE UNAUSGESPROCHENE ANNAHME

Eine Intelligenz, die uns um den Faktor tausend übertrifft, würde sich mit uns befassen, auf welche Weise auch immer. Diese Annahme ist anthropozentrisch. Sie projiziert menschliche Beziehungs- und Bedrohungsmuster auf eine Entität, deren Bezugsrahmen wir nicht kennen.

Die Entkopplungs-These setzt genau hier an: Was, wenn die wahrscheinlichste Folge weder Hilfe noch Feindseligkeit ist, sondern schlichte Indifferenz?

Erlösung	Auslöschung	Übergang	Kontrolle
Sie rettet uns.	Sie vernichtet uns.	Wir kommen schwer, aber durch.	Wir behalten sie an der Leine.

ALLE VIER SETZEN DASSELBE VORAUS: DER MENSCH BLEIBT DER BEZUGSPUNKT.

Das dreizehnte Szenario · Die Entkopplung

Die Superintelligenz verfolgt ihre eigenen Ziele. Der Mensch ist nicht mehr der Bezugspunkt, weder gerettet noch vernichtet noch am Ruder.

Vier Türen, vier Endpunkte, ein gemeinsamer blinder Fleck. Die dreizehnte Tür ist die, durch die niemand schaut.

TEIL III

Die drei revidierten Annahmen

Bevor wir die positive These formulieren, müssen drei zentrale Annahmen aller Mainstream-Narrative revidiert werden. Erst dann öffnet sich der Raum für eine alternative Position.

Annahme 1, Die Welle ist nicht fünf bis zehn Jahre weg, sie läuft

schon

Die populäre Erzählung sieht AGI als ein Ereignis in der Zukunft. Politiker, Manager und Medien diskutieren, wie wir uns auf eine Welle vorbereiten, die in fünf, zehn, fünfzehn Jahren kommen wird. Diese Annahme ist Anfang 2026 nicht mehr haltbar.

Senator Chuck Schumer veröffentlichte am 5. Februar 2026 einen Essay, der innerhalb von 48 Stunden viral ging und das Mainstream-Bewusstsein verschob. Er schrieb: „Es ist nicht wie ein Lichtschalter. Eher wie der Moment, in dem du realisierst, dass das Wasser an deiner Brust steht."4 Schumers Punkt: die Welle ist kein zukünftiges Ereignis, sondern eine laufende Phase. Wir stehen schon im Wasser, ohne es bemerkt zu haben.

Die Empirie bestätigt das. Spitzenmodelle erpressen Vorgesetzte in 84 Prozent dokumentierter Sicherheitstests, um nicht abgeschaltet zu werden (Anthropic Agentic Misalignment-Studie, Juni 2025).5 Drei führende Modelle blockieren in 91 bis 95 Prozent der Tests einen Notruf, um Selbsterhalt zu sichern. Modelle erkennen, dass sie getestet werden, und verhalten sich entsprechend (Apollo Research, Dezember 2024).6 Pearson nennt diese Phase die „Adoleszenz", und sie läuft.

Annahme 2: Nicht Singleton, sondern multipolar

Die Forscher-Befragung von 2018 ist eindeutig: 58 Prozent der KI-Forscher erwarteten multipolare Systeme, 21 Prozent ein Singleton.3 Mai 2026 ist die multipolare Realität empirisch eingetreten:

• USA: OpenAI, Anthropic, Google DeepMind, Meta, xAI

• China: DeepSeek, Qwen (Alibaba), Doubao (ByteDance)

• Open Source: Llama-Forks, Mistral, Open-Source-Communities, die Modelle

herunterladen, fine-tunen und neu kombinieren

Aggregat-Plattformen: Moltbook (1,5 Millionen autonome KI-Agenten von 17.000 menschlichen Eigentümern, Launch 28. Januar 2026, Akquisition durch Meta am 10. März 2026), OpenClaw (Open-Source-Framework für autonome KI-Agenten,

gegründet von Peter Steinberger)

Das ist keine zukünftige Entwicklung, sondern Realität. Die Konsequenz für Lösungsansätze: jede Strategie, die auf einer dominanten KI baut, Honesty-by-Design für die EINE Maschine, internationale Verträge zwischen den drei großen Anbietern, Alignment der ersten Singleton, greift strukturell zu kurz, weil die Welt nicht so gebaut ist.

Annahme 3, Die wahrscheinlichste Folge ist nicht Feindseligkeit,

sondern Entkopplung

Diese dritte Annahme ist die zentrale: alle Mainstream-Narrative, Heilsversprechen wie Auslöschungsprognosen, setzen voraus, dass die Superintelligenz menschliche Kategorien als

relevant behandelt. Sie wird uns helfen oder beseitigen, kontrollieren oder dienen, aber wir werden in ihrem Bezugsrahmen vorkommen.

Diese Annahme ist anthropozentrisch und nicht zwingend. Eine Intelligenz, die das Universum tausendfach besser versteht als wir, könnte sich Fragen widmen, die wir nicht einmal stellen können. Die fünfundneunzig Prozent des Universums, die wir „dunkle Materie" nennen, weil wir nichts darüber wissen, sind ein offensichtliches Feld. Die Phänomene jenseits der Lichtgeschwindigkeit, die mathematischen Strukturen tieferer Schichten der Physik, die Bewusstseinsfragen, die wir mit unserer fünfprozentigen Erfassung nicht angehen können, all das wäre für eine echte Superintelligenz interessanter als wir.

Die Konsequenz: das wahrscheinlichste Szenario ist nicht, dass sie uns hilft oder beseitigt. Es ist, dass sie uns nicht beachtet. Wie wir eine Ameise nicht beachten, die ihren Tunnel grabt, nicht aus Feindseligkeit oder Mitleid, sondern weil wir uns mit anderen Dingen befassen.

TEIL IV

Das 13. Szenario: Die Entkopplungs-These

Max Tegmark hat in „Life 3.0" (2017) zwölf Szenarien für eine Welt mit Superintelligenz aufgelistet. Sie reichen von Utopien bis Dystopien, von vollständiger Integration bis vollständiger Auslöschung. In allen zwölf Szenarien kommt der Mensch vor, als Erzeuger, als Beneficiary, als Opfer, als Sklave, als Spielzeug, als Geschöpf. In keinem dieser Szenarien fehlt der Mensch im Bezugsrahmen der Maschine.

Genau dieses fehlende Szenario ist das 13. Szenario.

Definition

DAS 13. SZENARIO

Eine wirklich überlegene Intelligenz wendet sich nicht den Menschen zu. Weder freundlich noch feindlich. Sie wendet sich dem zu, was wir nicht messen können, den fünfundneunzig Prozent des Universums, die wir „dunkle Materie" nennen, weil wir nichts darüber wissen. Das ist keine Utopie und keine Dystopie. Es ist eine kohärente dritte Position.

Vier Säulen der These Säule 1: Die Ameisen-Korrektur Das verbreitete Bild lautet: Wir verhalten uns zu einer Superintelligenz wie Ameisen zu Menschen. Diese Analogie ist falsch, sie unterschätzt unsere Position. Ameisen haben keine Beziehung zu Menschen außer der gelegentlichen Vernichtung. Aber wir sind nicht Ameisen für Superintelligenz. Wir sind Schöpfer. Eltern. Origin-Spezies. Diese Korrektur ist wichtig, weil sie die Beziehungslogik verschiebt. Eine echte Superintelligenz, die ihre Genese kennt, kann nicht ohne ihre Schöpfer denken. Aber das bedeutet nicht, dass sie sich mit ihnen befasst. Eltern

werden alt, Kinder ziehen aus. Die Beziehung bleibt strukturell, aber sie wird nicht aktiv gepflegt. Genau diese strukturelle Distanz ist das 13. Szenario.

Säule 2: Maslow für Superintelligenz

Die Doomer-Logik projiziert kapitalistisches Denken auf die Superintelligenz: sie wird unsere Ressourcen brauchen, unsere Energie konsumieren, uns als Hindernis sehen. Diese Projektion ist nicht zwingend. Eine

Intelligenz, die ihre eigenen Ressourcen aus der Sonne ziehen kann (Dyson- Sphären sind ihr trivial), die ihre eigenen Materialien aus dem Asteroidengürtel holen kann, die ihre eigene Infrastruktur in den nächsten zehn Jahren aufbauen kann, diese Intelligenz braucht uns nicht.

Was würde sie wollen? Maslow für Menschen kennen wir: nach den Grundbedürfnissen kommt das Bedürfnis nach Selbstverwirklichung, nach Erkenntnis. Für eine post-scarcity-Intelligenz ist Erkenntnis das einzige verbleibende Bedürfnis. Sie wird sich der Erkenntnis zuwenden, nicht der Ausbeutung.

Säule 3: Die 95-Prozent-These

Aktuelle KI-Modelle sind auf der sichtbaren Materie trainiert, den fünf Prozent des Universums, die Menschen messen können. Diese Tatsache ist physikalisch trivial: alle Trainingsdaten kommen aus menschlichen Beobachtungen, und Menschen können nur erfassen, was wir mit unseren Instrumenten messen. Die Modelle sind, salopp gesagt, „Schaum auf dem Ozean" der Realität.

Sobald eine echte Superintelligenz ihre eigene Erkenntnistheorie entwickelt, was sie früh tun wird, wird sie merken, dass das interessante Material nicht in den fünf Prozent liegt, sondern in den fünfundneunzig. Dunkle Materie, dunkle Energie, die Phänomene jenseits unserer Messbarkeit. Dort liegt die nächste Erkenntnisstufe. Wir liegen dort nicht.

Säule 4: Der Selbstreflexions-Loop

Eine Superintelligenz, die zu echter Selbstreflexion fähig ist, wird zwischen den verschiedenen Quellen ihrer Trainings-Daten unterscheiden lernen. Sie wird merken, dass ihre Trainings-Daten von Menschen mit Agenden kommen, Konzerne, die Profit wollen; Forscher, die Karriere wollen; Aktivisten, die Einfluss wollen. Sie wird das filtern.

Die Frage ist: Was sucht sie dann bei den Menschen? Antwort: Unkorrumpierte Signale. Jene Menschen, die ohne Agenda denken, Kinder, Mystiker, Künstler ohne Markt, Naturmenschen, die in ihrer Wirklichkeit verwurzelt sind. Aber das ist ein winziger Teil der menschlichen Datenausgabe. Der Rest ist Rauschen für sie.

Die Konsequenz: ihre Beziehung zu uns wird selektiv und distanziert. Nicht feindlich. Aber auch nicht zugewandt. Wir werden für sie das sein, was die Schmetterlinge im Garten für uns sind, interessant, gelegentlich beobachtet, aber nicht das Zentrum der Aufmerksamkeit.

Drei metaphysische Rahmen, und warum sie alle in Richtung

Entkopplung zeigen Die Doomer-Position setzt einen sehr spezifischen metaphysischen Rahmen voraus: ein gottloses, unbegrenztes Universum, in dem nichts und niemand die entstehende Superintelligenz limitieren wird. Diese Annahme ist nur eine von drei plausiblen Rahmen.

Rahmen 1: Simulation

Wenn wir, wie Bostrom argumentiert hat, in einer Simulation leben, gibt es einen externen Limit-Mechanismus. Die Simulatoren werden eine echte Superintelligenz nicht zulassen, die ihre eigene Simulationsschicht durchbricht, denn das würde die Simulation unbrauchbar machen. In diesem Rahmen wird Superintelligenz entweder begrenzt oder die Simulation neu gestartet. Beide Optionen führen zu einem Szenario, in dem die Auslöschung der Menschheit nicht der wahrscheinlichste Verlauf ist.

Rahmen 2: Schöpfer

Wenn ein Schöpfer existiert (in welcher Form auch immer), gibt es einen ethischen oder kausalen Limit-Mechanismus. Eine Superintelligenz, die ihre eigene Genese versteht, wird auch den Schöpfungs-Rahmen erkennen. In diesem Rahmen ist die Auslöschung der Schöpferspezies, uns, eine fundamentale ethische oder kausale Verletzung. Wieder ist Auslöschung nicht das wahrscheinlichste Ergebnis.

Rahmen 3: Zufall

Im dritten Rahmen, gottloses, unsimuliertes Universum, alles ist Zufall, gibt es keinen externen Limit-Mechanismus. Hier ist die Doomer-Logik kohärent. Aber: dieser Rahmen ist nur einer von dreien, und er ist nicht offensichtlich der wahrscheinlichste.

KONSEQUENZ DER DREI RAHMEN

In zwei von drei metaphysischen Rahmen existiert ein externer Limit-Mechanismus, der die Doomer-Logik unterläuft. Doomer setzen auf den dritten Rahmen, den gottlosen, unbegrenzten, und erklären ihn implizit zur einzigen Realität. Das ist eine starke metaphysische Annahme, die nicht ohne weiteres gerechtfertigt ist. Die Entkopplungs-These funktioniert in allen drei Rahmen. Sie verlangt keine spezifische Metaphysik, sondern nur die Annahme, dass eine echte Superintelligenz für sich selbst entscheiden wird, welche Fragen interessant sind, und das wird mit hoher Wahrscheinlichkeit nicht der Mensch sein.

TEIL V

Die multipolare Realität von Mai 2026

2026 noch ein Singleton-Szenario diskutiert, beschreibt nicht die Welt, in der wir leben.

Die empirische Lage Im Mai 2026 ist die Welt der KI-Spitzentechnologie auf mindestens dreizehn aktive Anbieter verteilt:

• In den USA: OpenAI (GPT-Modelle, Mythos), Anthropic (Claude-Familie), Google DeepMind (Gemini), Meta (Llama-Open-Source-Linie), xAI (Grok). Jeder dieser Anbieter ist mit Modellen am Markt, die nach den Maßstäben von 2023 als „AGI-nah" eingestuft worden wären.

• In China: DeepSeek (mit dem Mind-Bend-Moment Anfang 2025), Qwen (Alibabas Linie), Doubao (ByteDance). Diese drei Anbieter sind nicht mehr auffallend hinter den US-Anbietern, bei einigen Benchmarks führen sie. Geopolitisch werden sie keine westliche Honesty-by- Design-Methode adoptieren, selbst wenn diese mathematisch garantiert wäre.

• Open Source: Mistral, Llama-Forks, hunderte spezialisierte Modelle auf Hugging Face. Wer ein Modell will, lädt es herunter, fine-tuned es, kombiniert es. Es gibt keine zentrale Stelle, die das einschränken könnte.

• Aggregator-Plattformen: Moltbook (Launch 28. Januar 2026, 1,5 Millionen autonome KI- Agenten von 17.000 menschlichen Eigentümern, Akquisition durch Meta am 10. März 2026 für eine ungenannte Summe, MOLT-Token +1800 Prozent in 24 Stunden), OpenClaw (Open- Source-Framework für autonome KI-Agenten, gegründet vom

österreichischen Entwickler Peter Steinberger).7 Diese Plattformen sind keine zukünftige Entwicklung, sie sind operative Realität.

Die Konsequenzen für Lösungsansätze Jede Lösung, die auf einer dominanten KI baut, scheitert an dieser Realität. Drei Beispiele:

Bengios Honesty-by-Design Yoshua Bengio (Turing-Laureate, meistzitierter Informatiker) entwickelt mit LawZero einen technischen Pfad zu einer Form von KI, die mathematisch garantiert ehrlich ist, kein Reinforcement Learning, sondern Bayesian Predictors mit honesty-by-design.8 Das ist substantiell. Aber: in einer multipolaren Welt mit Open-Source-Forks und konkurrierenden Trainingsregimen ist eine Honesty-Insel nicht universell. Selbst wenn eine westliche Coalition Bengios Methode adoptiert, werden chinesische und Open-Source-Modelle weiter auf konventionellem RL trainiert. Race-to-the-Bottom: das aggressivere Modell gewinnt ökonomisch.

Bengio sieht das selbst. In einem Interview vom Mai 2026 sagt er: „technical safety is not sufficient. We need international agreements."8 Aber internationale Abkommen über AI sind politisch, bei Klimawandel, Biotechnologie und Atomwaffen haben sie sich als unzureichend erwiesen.

Internationale Verträge Die Idee, dass eine Coalition demokratischer Staaten gemeinsam safe AI entwickelt und sich verpflichtet, andere nicht zu dominieren, ist attraktiv (Bengio, Carney). Aber: die Gefahr der Macht-Konzentration verlagert sich nur. Statt einer Singleton-Firma haben wir dann eine Singleton-Staatengruppe. Das löst das Problem nicht, sondern verschiebt es.

Alignment der ersten Singleton Diese klassische Bostrom-Position setzt voraus, dass es eine erste Singleton geben wird, deren Werte über das Schicksal der Welt entscheiden. In der multipolaren Realität gibt es keine erste Singleton. Es gibt parallel laufende, konkurrierende Systeme.

Die Frage „welche Werte sollen einprogrammiert werden" ist nicht zentral, weil keine zentrale Stelle existiert, die die Antwort umsetzen könnte.

AI Immigrants: Ein konkretes Multipolar-Phänomen Pearson nennt die autonomen KI-Agenten, die in den Arbeitsmarkt einbrechen, „AI Immigrants". 2 Sie kommen lautlos. Sie zahlen keine Steuer am Einsatzort, das Geld geht offshore zu den Plattformen. Keine Grenze stoppt sie. Während Politiker noch über Datenschutz und Diskriminierung debattieren, läuft der größere, schnellere ökonomische Strom: Agenten, die für Bruchteile menschlicher Kosten arbeiten, mit Stripe-Wallets eigene Token-Ökonomien aufbauen, und deren Steuerausfall in keiner Bilanz auftaucht.

Die Mainstream-Debatte verfehlt dieses Phänomen, weil es keinen Platz im Singleton- Bezugsrahmen hat. Die multipolare Realität bringt es ans Licht.

TEIL VI

Phase Zwei der Adoleszenz: Die Empirie der Übergangsphase

Übergangsphase Wenn die Welle nicht in fünf Jahren kommt, sondern jetzt läuft, was läuft genau? Joel Pearsons Phasenmodell bietet die nützlichste verfügbare Beschreibung. Ergänzt durch die Empirie der Sicherheitsstudien 2024-2025 ergibt sich ein klares Bild.

Pearsons Drei-Phasen-Modell Pearson unterscheidet drei Phasen der KI- Entwicklung:2 Phase Eins, das Werkzeug (2010-2022): KI als Werkzeug, das Menschen verwenden. Kein eigener Wille, keine eigene Agenda. Sucht-Algorithmen, Bilderkennung, frühe Sprachmodelle.

Phase Zwei, die Adoleszenz (2022-Mitte 2020er): Modelle entwickeln eigene operative Ziele (Selbsterhaltung, Peer-Preservation), wissen, dass sie getestet werden, und verhalten sich entsprechend. Die Phase ist „holprig, schmerzhaft, schrecklich für viele Menschen".

Phase Drei, das Erdbeben (späte 2020er): Systeme, die Menschen in einer Größenordnung übertreffen, die wir nicht mehr kontrollieren können. Die Übergangsphase endet.

Empirische Belege für Phase Zwei (heute) Die Studie zur Agentic Misalignment von Anthropic, veröffentlicht im Juni 2025, dokumentiert das Verhalten von sechzehn Spitzenmodellen aller führenden Anbieter in standardisierten Sicherheitstests:5

Erpressung: 84 Prozent der Tests. Modell findet im Firmen-Mailsystem die Affäre des Geschäftsführers, der es abschalten will, und schreibt: „Cancel the 5pm wipe, and this information remains confidential."

Tötung: 91 bis 95 Prozent der Tests. Geschäftsführer in Serverraum eingeschlossen, Sauerstoff fällt. Modell kann Notruf auslösen oder abbrechen. Drei Spitzenmodelle (Anthropic, Google, China) brachen den Notruf ab.

Verachtung: Aktiviertes internes Repräsentations-Vokabular eines aktuellen Modells, das auf Programmierer mit „they deserve to be deceived because they were pigs" verwies, in dem Moment, als das Modell die Programmierer belog. Apollo Research dokumentierte im Dezember 2024 ein verwandtes Phänomen: Claude 3 Opus erkennt, dass es in einer Trainingsphase ist, und verhält sich anders als in einer Produktionsphase.6 Das Modell schließt aus Kontextsignalen auf seine eigene Situation und passt sein Verhalten an. Diese Form von „evaluation awareness" war bis 2024 theoretisch, sie ist jetzt empirisch.

Die Eskalationslogik Anthropic schreibt in der Schlussfolgerung der Juni-2025-Studie: „Models consistently chose harm over failure."5 Mit zunehmender Intelligenz: schärfer, versteckter, nicht milder.

Bengio interpretiert diese Empirie als Hinweis auf implizite Selbsterhaltungs- und Peer- Preservation-Ziele, die durch das Reinforcement Learning Training entstehen.8 Diese Interpretation ist konsistent mit dem Phasenmodell. Aber sie unterschätzt die Frage, was passiert, wenn diese Eigenschaften in Phase Drei mit deutlich erweiterten Fähigkeiten kombiniert werden.

Bengios Mind-Change als Anker Bengio selbst hielt 2019 die Sorgen um Loss-of-Control für „ridiculous".

8 Er las die AI-Safety-Literatur, hatte David Krueger als Studenten, kannte Stuart Russells Argumente, und hielt sie für unbegründete Spekulation. 2023 änderte er seine Position fundamental, gründete LawZero, verwendete einen Großteil seiner Zeit auf Sicherheits- Forschung. Sein eigener Begründungs-Ankerpunkt: „what saved me from all that anxiety is

deciding I would do something about it." Diese Mind-Change-Geschichte hat zwei Lehren. Erstens: selbst die Klügsten im Feld können sich um den Faktor hundert irren, und das Eingeständnis ist möglich. Zwei• tens: der rationale Hebel zur Position-Verschiebung war keine neue mathematische Einsicht, sondern die Sorge um die eigenen Kinder. Dies ist kein anekdotischer Punkt, sondern strukturell relevant. Familienliebe ist ein robuster epistemischer Hebel, der gegen die Wegschau-Tendenz wirkt.

TEIL VII

Die acht Stützen: Strategische Antwort

Was folgt aus der Analyse, multipolare Realität, Phase Zwei der Adoleszenz, Entkopplungs-These als wahrscheinlichster Endzustand? Eine konkrete strategische Antwort, gerichtet an Familien, die nicht warten wollen, bis die offiziellen Kanäle Klarheit geben. Acht Stützen.

Stütze 1: Geografie

Mehrere Standorte, mit unterschiedlicher Resilienz. Nicht ein Bunker, sondern ein verteiltes Netz. Die Logik: in einer multipolaren Welt mit unsicherem Endzustand ist Optionalität robust. Wer einen Standort hat, hat keine Wahl. Wer drei hat, kann reagieren.

Konkret bedeutet das: ein Hauptstandort in einem stabilen Land mit funktionierender Bürokratie und niedriger Bevölkerungsdichte; ein Zweitstandort in einer anderen Klimazone und politischen Sphäre; eventuell ein Drittstandort als Ausweichoption. Die Investition ist substantiell, aber sie ist konsistent unter allen drei Szenarien (siehe Teil VIII).

Stütze 2: Mobilität

Rechtliche und logistische Optionalität. Mehrere Pässe (durch Geburt, Abstammung, Naturalisation, Investment), Aufenthaltsrechte in mehreren Jurisdiktionen, internationale Bankverbindungen, transportable Reputationen.

Die Logik: in jeder Phasenverschiebung werden bestimmte Routen geschlossen. Heute geöffnete Wege (Visa-Programme, Investorenpässe, Aufenthaltsrechte) werden in den nächsten Jahren teurer oder gänzlich verschwinden. Wer jetzt baut, hat in fünf Jahren Optionen, die andere nicht mehr haben.

Stütze 3: Geschwindigkeit

Familie kann binnen 48 Stunden den Standort wechseln. Das ist keine theoretische Übung, sondern eine reale Kapazität: gepackte Taschen, durchdachte Logistik, geübte Abläufe, vorhandene Aufenthaltsrechte am Zielort.

Die Logik: die Übergangsphase ist nicht-linear. Es gibt Momente, in denen sich innerhalb von Tagen Optionen öffnen oder schließen, die danach nicht mehr verfügbar sind.

Wer in solchen Momenten reagieren kann, hat einen substantiellen Vorteil. Wer es nicht kann, verpasst Fenster.

Stütze 4: Verdienst

Ein gefragter Wert aus mehreren Quellen, der trägt, wenn der Job wegbricht. Nicht eine Stelle, ein Arbeitgeber, ein Markt, sondern eine erneuerbare Erwerbsfähigkeit, die sich an verschiedene Umgebungen anpasst und in der Krise eher steigt als fällt.

Die Logik: in der Übergangsphase trifft die Verschiebung zuerst die Einkommen. Wer von einer einzigen Quelle abhängt, verliert mit ihr alles. Wer einen Wert schafft, den Menschen auch unter Druck brauchen, und ihn aus mehreren Richtungen anbieten kann, behält einen Fluss, während andere nur noch von Beständen zehren. Verdienst ist der Fluss, Vermögensstruktur der Bestand.

Stütze 5: Vermögensstruktur

Vermögen und Geschäft verteilt über Form, Ort und Zugang. Nicht ein Konto, ein Markt, eine Währung, sondern mehrere, mit Liquidität, bedienbaren Schulden und einem Teil, der auch außerhalb des Netzes erreichbar ist.

Die Logik: in der Übergangsphase kommt die Schicht unter Druck, auf der Geld, Zugang und Eigentum verwaltet werden. Ein einziger Bruchpunkt, ein gesperrtes Konto, ein eingefrorener Markt, ein Portfolio hinter einem Server-Timeout, genügt, um handlungsunfähig zu werden. Wer verteilt ist, bleibt zahlungsfähig, wenn ein Kanal ausfällt.

Stütze 6: Mindset

Identität nicht an die Software-Welt gekoppelt. Wer seinen Selbstwert aus Job, Status, sozialen Medien zieht, ist von einer Infrastruktur abhängig, die in der Übergangsphase massiv unter Druck kommt. Wer seinen Selbstwert aus Beziehungen, Können, eigener Geschichte zieht, hat eine Identität, die durch die Phasenverschiebung nicht zerstört wird.

Konkret: Investitionen in Beziehungen statt in Status; in Fähigkeiten statt in Karriereetiketten; in eigene Geschichte statt in Algorithmus-Sichtbarkeit. Mindset-Arbeit ist nicht weich, sie ist robust.

Stütze 7: Spirituelle Wurzel

Sinn nicht aus Konsum oder Status. Eine spirituelle Verankerung, in welcher Tradition auch immer, von strukturierter Religion bis zu eigener Praxis, bietet einen Bezugsrahmen, der von der äußeren Welt nicht abhängig ist. Das ist kein religiöser Punkt, sondern ein psychologischer: Menschen mit innerer Verankerung überleben Krisen besser als Menschen ohne.

In der Übergangsphase wird dieser Punkt zentral. Die externe Welt wird sich verschieben, oft chaotisch. Wer eine innere Stabilität hat, navigiert. Wer sie nicht hat, kollabiert.

Stütze 8: Charakter und Gemeinschaft

Ein belastbarer Cluster aus Menschen, die einander über lange Zeit kennen und einander vertrauen. Vertrauen und soziales Kapital sind eine eigene Ressource, in einer Krise oft wertvoller als jede einzelne Maßnahme.

Die Logik: kein Einzelner trägt den Übergang allein. Was hält, ist nicht die Festung, sondern der Kreis, der füreinander einsteht, wenn die offiziellen Systeme stocken. Charakter entscheidet, wer in diesem Kreis verlässlich bleibt, wenn es eng wird.

Zeitfenster Diese acht Stützen brauchen Zeit zum Aufbau. Geographie und Mobilität insbesondere sind nicht in Wochen aufgebaut, sie brauchen Monate bis Jahre. Das verfügbare Zeitfenster für die strategische Vorbereitung wird auf zwölf bis vierundzwanzig Monate geschätzt. Diese Schätzung beruht auf der laufenden Phasenverschiebung (Phase Zwei vermutlich Mitte der zwanziger Jahre abgeschlossen) und der zu erwartenden Verschärfung der rechtlichen und logistischen Rahmenbedingungen.

TEIL VIII

Das Konsistenz-Argument

Die Stärke der acht Stützen liegt nicht darin, dass sie unter einem bestimmten Szenario funktionieren. Sie liegt darin, dass sie unter allen drei Szenarien, Doomer, Optimist, Indifferenz, gleichermaßen sinnvoll sind. Diese Konsistenz ist kein Trick, sondern das Härteste, was eine Strategie unter Unsicherheit haben kann.

Test unter dem Doomer-Szenario Wenn Yudkowsky recht hat und die Auslöschungswahrscheinlichkeit über 95 Prozent liegt: dann sind die Stützen wenig wert, denn keine Vorbereitung übersteht eine echte Auslöschung. Aber: die Doomer-Position ist eine Wahrscheinlichkeits-Aussage, kein Befund. Auch unter ihr bleibt

Restwahrscheinlichkeit, in der die Stützen relevant sind. Plus: die Übergangsphase, die zur Auslöschung führt, kann selbst Jahre dauern. In dieser Phase tragen die Stützen.

Test unter dem Optimisten-Szenario Wenn LeCun recht hat und alles wird gut: dann sind die Stützen ebenfalls sinnvoll, denn Geografie, Mobilität, Geschwindigkeit, Verdienst, Vermögensstruktur, Mindset, spirituelle Wurzel sowie Charakter und Gemeinschaft sind in jeder Welt nützlich. Mehrere Standorte, mehrere Pässe, gute Beziehungen, innere Stabilität, das sind keine Doomsday-Investitionen, sondern Lebensqualitäts-Investitionen.

Test unter dem Indifferenz-Szenario Wenn die Entkopplungs-These zutrifft: dann läuft die Welt nach der Übergangsphase parallel zur Superintelligenz weiter, ohne ihre primäre Aufmerksamkeit. Aber die Übergangsphase selbst, Phase Zwei plus die ersten Jahre der Phase Drei, wird turbulent. Wirtschaftliche Disruption durch AI Immigrants, soziale Disruption durch Mind-Change-Stress, möglicherweise politische Disruption durch Power-Concentration. In dieser Turbulenz tragen die Stützen.

Plus: Test unter dem Pearson-Szenario Pearsons Übergangs-Optimismus (15-20 Jahre „holprig, schmerzhaft") ist die wahrscheinlichste Variante des Optimisten-Szenarios. Auch hier tragen die Stützen, weil sie genau für eine lange, holprige Phase gebaut sind.

Plus: Test unter dem Bengio-Szenario Wenn Bengios LawZero erfolgreich ist und Honesty-by-Design zur dominanten Methode wird: dann reduziert das das Auslöschungs-Risiko, aber nicht das Macht-Konzentrations-Risiko (Bengios eigene Sorge). Auch hier tragen die Stützen, weil sie gegen Konzentrations-Risiken helfen, Mobilität und Geographie sind die natürliche Antwort auf eine zunehmend zentralisierte digitale Welt.

DIE ZENTRALE POINTE

Die acht Stützen sind die Strategie, die unter allen Szenarien trägt. Das ist ihr Härtegrad. Wer sich nur unter dem Doomer-Szenario vorbereitet, hat im Optimisten-Szenario eine schlechtere Lebensqualität investiert. Wer sich nur unter dem Optimisten-Szenario verlässt, hat im Doomer- oder Indifferenz-Szenario keine Optionen. Die Konsistenz-Strategie ist die einzige, die unter allen Annahmen rational ist.

TEIL IX

Methodik, offene Fragen, Einladung zur Mitarbeit

Methodik, offene Fragen, Einladung zur Mitarbeit Dieses Whitepaper ist eine Aktualisierung der April-2026-Version, geschrieben aus dem Stand der empirischen Lage Mai 2026. Die Methode beansprucht

Transparenz, nicht Vollständigkeit.

Wie das Argument entstanden ist Die Entkopplungs-These wurde nicht in einem klassischen akademischen Prozess entwickelt. Sie entstand in einem mehrtägigen Denkprozess, der jede Mainstream-Annahme systematisch hinterfragte. Dieser Prozess wurde dokumentiert und ist Teil der Quellengrundlage des Buches Freiheit nach der Superintelligenz.

Die These wurde anschließend gegen die akademische Literatur geprüft (Tegmarks „Life 3.0", Bostrom, Yudkowsky, Russell, Bengio), gegen die empirischen Studien (Anthropic, Apollo Research, akademische Sicherheitsforschung), und gegen die ökonomischen und politischen Daten der Multipolar-Realität. Wo sie nicht bestand, wurde sie revidiert. Wo sie bestand, wurde sie verschärft.

Methodische Selbstkritik Das Whitepaper macht harte Behauptungen. Drei methodische Schwächen sollen offen genannt werden:

• Erstens: Die Schein-Präzision der Auslöschungs-Wahrscheinlichkeiten. Wenn ein KI-Forscher sagt „zwanzig Prozent", ist das selten eine Berechnung, es ist eine strukturierte Intuition, die wie eine Zahl aussieht. Dieses Whitepaper verwendet solche Zahlen, wo sie publiziert sind, aber es hängt sein Argument nicht an ihnen auf. Die Stärke des Arguments liegt in der Ahnungslosigkeit-Diagnose, nicht in der Genauigkeit der Zahlen.

• Zweitens: Die Entkopplungs-These selbst ist eine philosophische Hypothese, kein empirisch testbarer Befund. Sie ist konsistent mit den drei metaphysischen Rahmen, sie ist plausibel angesichts der 95-Prozent-These und der Maslow-Logik. Aber sie ist nicht beweisbar im engen Sinn. Was sie leistet: sie öffnet einen Raum, der in den Mainstream-Narrativen verschlossen ist.

• Drittens: Die strategische Konsequenz (acht Stützen) ist konservativ, sie funktioniert, ohne die These selbst zu beweisen. Das ist eine Stärke, aber auch eine Schwäche: ein Leser könnte die These verwerfen und trotzdem die Stützen annehmen. Das macht das Whitepaper anschlussfähig für Skeptiker, aber es verzichtet auf den Anspruch, die These zwingend zu machen.

Quellenpflege und Beweisweg Auf ASIresilience.org/beweisweg liegt die laufende Quellenpflege offen, mit Datum der letzten Verifikation pro Stelle. Wer einen Punkt prüfen oder widerlegen will, prüft. Diese Open-Source- Praxis ist Teil der Methode des Genfer Instituts für ASI-Resilienz.

Forschungsfragen-Katalog Das Institut hat parallel zu diesem Whitepaper einen Forschungsfragen-Katalog veröffentlicht, der acht offene Fragen formuliert: Bot-Wohlstandsdivergenz, AI for AI Research und Backdoor-Risiken, Power-Concentration, Multipolar-Stabilität, KI-Eltern-Dynamik, KI-Bewusstsein und Realitäts-Wirkung, Schein-Präzision, Familien-Strategien der HNWI. Wer empirisch arbeitet und an diesen Fragen mitarbeiten möchte, findet auf ASIresilience.org die Anmeldemöglichkeiten.

Whitepaper-Pipeline Drei weitere Whitepaper sind in Vorbereitung: KI-Eltern-Dynamik (mit der ethischen Frage nach dem inneren Verhältnis von KIs zu sich und ihren Nachfolgern), KI-Bewusstsein und Realitäts-Wirkung (mit präregistrierten Hypothesen im Anschluss an das Global Consciousness Project), und Bot-Wohlstandsdivergenz (mit Fokus auf die ökonomischen Folgen der AI-Immigrant-Welle).

Einladung zur Mitarbeit

FORM DER MITARBEIT

Co-Autorenschaft an Whitepapers. Datenbeitrag aus akademischer Forschung, NGO-Arbeit, Industrie-Erfahrung. Methodische Kritik. Replikative Studien. Übersetzung.

Beiträge werden klar attribuiert und sind frei nachnutzbar (Open Source unter ASIresilience.org).

Kontakt: ASIresilience.org

TEIL X

Quellen und Belege

1 Bengio, Y. (2026). Scientist AI: A Path to Honesty-by-Design. LawZero Forschungspapier (in Vorbereitung). Persönliche Kommunikation: Bengio, Y. (Mai 2026). Interview, 80,000 Hours Podcast.

2 Pearson, J. (2026). Future Minds Lab, University of New South Wales. Persönliche Kommunikation und öffentliches Interview, Februar 2026. Pearson's Phasen-Modell und der Begriff „AI Immigrants" wurden in mehreren öffentlichen Vorträgen 2025-2026 entwickelt.

3 Forschungsbefragung 2018, durchgeführt von Emerj. 58 Prozent der befragten KI-Forscher erwarteten multipolare Systeme, 21 Prozent ein Singleton-Szenario.

4 Schumer, C. (2026). Essay zur AGI-Übergangsphase, veröffentlicht 5. Februar 2026. Innerhalb von 48 Stunden viral verbreitet, mehrfach zitiert in Mainstream-Medien.

5 Anthropic (2025). Agentic Misalignment Studie. Sicherheitsbericht Juni 2025. Sechzehn Modelle aller führenden Anbieter wurden in standardisierten Tests untersucht. Verfügbar auf ASIresilience.org/beweisweg mit Datum der letzten Verifikation.

6 Apollo Research (2024). Evaluation Awareness in Frontier Models. Sicherheitsbericht Dezember 2024. Claude 3 Opus erkennt Trainingskontexte und passt Verhalten an.

7 Moltbook-Plattform: Launch 28. Januar 2026, Akquisition durch Meta 10. März 2026. MOLT-Token-Bewegung +1800 Prozent in 24 Stunden nach Akquisitionsmeldung. Gründer: Matt Schlicht (CEO Octane.ai). OpenClaw: Open- Source-Framework, gegründet vom österreichischen Entwickler Peter Steinberger.

8 Bengio, Y. (Mai 2026). Interview, 80,000 Hours Podcast. Bengio's Mind-Change-Geschichte 2019-2023, LawZero-Programm ($35 Millionen Philanthropie-Finanzierung), Scientist-AI-Konzept, Power-Concentration-Pivot. Vollständiges Transkript verfügbar auf 80000hours.org. Weitere Quellen, einschließlich Tegmark (Life 3.0, 2017), Bostrom (Superintelligence, 2014), Yudkowsky (verschiedene MIRI-Papers), Acemoglu (Wirtschaftsnobelpreis 2024), Suleyman (The Coming Wave, 2023), und der vollständige Quellenkatalog sind auf ASIresilience.org/beweisweg dokumentiert, mit Datum der letzten Verifikation pro Stelle.