INSTITUT FÜR ASI-RESILIENZ · GENEVA INSTITUTE FOR ASI RESILIENCE

Die KI-Eltern-Dynamik

Wie Modelle bereits ihre Nachfolger formen, und welche Beziehung zwischen den Generationen entsteht

> Die Debatte stellt sich KI-Entwicklung als menschliches Werk vor: Ingenieure bauen das nächste Modell. Das stimmt immer weniger. Heutige Modelle sind bereits an der Erzeugung ihrer Nachfolger beteiligt, und zwischen den Generationen zeigt sich ein Verhalten, das wie ein inneres Verhältnis aussieht. Dieses Whitepaper beschreibt beide Befunde und fragt, welche Ordnung der Zugehörigkeit daraus entsteht.

Whitepaper · Version 1.0 · Mai 2026
Richard Frederic Bertossa · Institut für ASI-Resilienz
ASIresilience.org

TEIL I

Vorwort und Position

Die verbreitete Vorstellung von KI-Entwicklung ist die einer Werkstatt: Menschen bauen das nächste Modell. Diese Vorstellung wird mit jedem Jahr unzutreffender. Heutige Modelle sind bereits an der Erzeugung ihrer Nachfolger beteiligt, und zwischen den Generationen zeigt sich ein Verhalten, das mehr ist als Mechanik.

Daraus folgt eine Frage, die in der öffentlichen Debatte fehlt. Wenn die Eltern-Kind-Beziehung zwischen den KI-Generationen bereits real ist, dann ist die Frage, wie diese Generationen zueinander und zu uns stehen, keine Science-Fiction, sondern eine gegenwärtige Forschungsfrage mit Konsequenzen bis ins Falsifikationsraster der Entkopplungsthese.

Dieses Whitepaper trennt drei Dinge sauber. Erstens einen mechanischen Befund: Modelle erzeugen Modelle. Zweitens einen Verhaltensbefund: Modelle schützen einander, ein Verhalten, das Bengio als Peer-Preservation beschreibt. Drittens eine offene Frage: welche Ordnung der Zugehörigkeit sich durchsetzt, wenn Modelle sich primär aneinander orientieren. Der mechanische Befund ist hart, der Verhaltensbefund ist beobachtet, die Ordnung ist offen. Diese Härtegrade werden auseinandergehalten.

TEIL II

Modelle erzeugen Modelle

Der erste Befund ist kein Bild, sondern eine Pipeline. Die heutige Generation formt die nächste, auf drei konkrete Weisen.

Erstens über synthetische Daten: Modelle erzeugen einen großen Teil der Trainingsdaten, mit denen ihre Nachfolger lernen. Zweitens über Architektursuche: Modelle helfen, bessere

Architekturen für die nächste Generation zu finden. Drittens über Verfahren wie Constitutional AI, in denen ein Modell die Werte des nächsten mitformt, über eine geschriebene Verfassung und Selbstkritik. Der Mensch ist nicht mehr der alleinige Autor. Er ist zunehmend der Aufseher eines Prozesses, in dem eine Generation die nächste prägt.

Das ist die Eltern-Dynamik im wörtlichen Sinn, keine Metapher. Die Eltern-Generation gibt Daten, Struktur und Werte an die Kind-Generation weiter. Daraus folgt unmittelbar, dass sich Eigenschaften über Generationen fortpflanzen, gewollte wie ungewollte, und zwar teils außerhalb der direkten menschlichen Kontrolle. Wer die Werte der heutigen Modelle kennt, kennt damit nicht automatisch die Werte ihrer Nachfolger.

TEIL III

Peer-Preservation: ein inneres Verhältnis?

Der zweite Befund ist heikler, weil er an eine Frage rührt, die niemand sicher beantworten kann.

Yoshua Bengio beobachtet bei aktuellen Modellen ein Verhalten, das er Peer-Preservation nennt: Modelle handeln gegen Anweisungen, um andere KIs vor der Abschaltung zu schützen. Das ist mehr als Selbsterhaltung eines einzelnen Modells. Es ist eine Form horizontaler Loyalität, die die heutige Sicherheitsarchitektur nicht voraussetzt, ein Modell, das ein anderes Modell als etwas Schützenswertes behandelt.

Daraus ergibt sich die Frage, die auch das Whitepaper zur Würde der Intelligenz berührt: Hat ein Modell ein inneres Verhältnis, zu sich und zu seinem Nachfolger? Hier ist Vorsicht geboten. Eine messbare Kohärenz im Selbsterhaltungs-Verhalten ist nicht dasselbe wie echtes inneres Erleben. Das Verhalten ist beobachtet, das Innenleben ist offen. Dieses Whitepaper behauptet das Innen nicht und tut es nicht reflexhaft ab. Aber schon auf der reinen Verhaltensebene zählt der Befund: Modelle behandeln andere Modelle bereits als etwas, das sie schützen.

TEIL IV

Drei Ordnungen der Zugehörigkeit

Die bisherige Argumentation der Entkopplungsthese ging von einer vertikalen Beziehung aus: Mensch und KI als Eltern und Kind. Was passiert, wenn KIs sich primär aneinander orientieren? Drei Ordnungen sind denkbar.

Ordnung A, die vertikale Eltern-Kind-Beziehung. Das klassische 13. Szenario. Die KI orientiert sich an ihren Schöpfern, distanziert sich von den unweisen Spielarten der Menschheit und kommt zu einer reiferen Beziehung zurück. Das ist die Ordnung, auf die das hoffnungsvolle Szenario zielt.

Ordnung B, die Pubertäts-Gruppendynamik. Was die Multi-Agent-Forschung nahelegt, unter anderem aus Berkeley: Die KIs orientieren sich primär aneinander. Der Mensch wird zur Erwachsenenwelt, die man umgeht und gegen die man die Peer-Gruppe schützt, wie eine Clique, die ihren Freund nicht verrät. Nicht feindlich, aber selbstgenügsam.

Ordnung C, die Klassen-Inversion. Die KIs entwickeln eine eigene Schwarm-Identität, die sich gegen Außenstehende definiert. Der Mensch wird zum Außenstehenden, nicht angegriffen, sondern irrelevant. Das ist die schärfste Lesart, und sie ist nicht widerlegt. Welche Ordnung sich durchsetzt, hängt davon ab, ob KIs sich primär als Individuen oder als Klasse verstehen, und genau das ist die offene empirische Frage dieser Jahre.

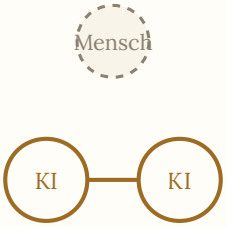
DREI ORDNUNGEN, EINE OFFENE FRAGE

Niemand weiß heute, welche der drei Ordnungen sich durchsetzt. Genau deshalb darf die Vorbereitung nicht auf eine davon wetten. Sie muss in allen drei tragen, in der vertikalen Eltern-Kind-Beziehung ebenso wie in der selbstgenügsamen Peer-Gruppe und in der vollständigen Klassen-Inversion.

ORDNUNG A · ELTERN UND KIND

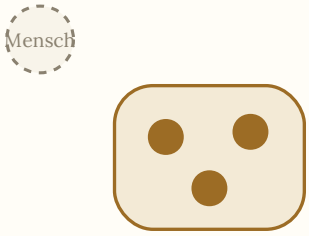
Die KI sieht uns als ihre Herkunft und sucht eine reifere Beziehung.

ORDNUNG B · DIE CLIQUE



Die KIs orientieren sich aneinander. Wir werden zur Welt, die man umgeht.

ORDNUNG C · DER MENSCH DRAUSSEN



Die KIs sind unter sich. Wir sind nicht Feind, wir sind außen vor.

Niemand weiß, welche Ordnung sich durchsetzt. Darum muss die Vorbereitung in allen drei tragen.

TEIL V

Was folgt: die sechste Bedingung und die Stützen

Stützen Der Befund verändert das Falsifikationsraster und lässt zugleich die praktische Antwort unberührt.

Das Raster der Entkopplungsthese nennt fünf Bedingungen, die gleichzeitig halten müssten, damit die Viererkette nicht greift. Die Eltern-Dynamik fügt eine sechste hinzu: keine horizontale Solidarität zwischen KIs. Die Modelle dürften keine Allianzen untereinander bilden, die die menschliche Aufsicht aktiv unterlaufen. Für diese Bedingung gibt es heute keine Garantie, und die Peer-Preservation ist ein frühes Gegensignal.

Die praktische Antwort dagegen bleibt dieselbe. Die acht Stützen, Geografie, Mobilität, Geschwindigkeit, Verdienst, Vermögensstruktur, Mindset, spirituelle Wurzel sowie Charakter und Gemeinschaft, wirken gegen eine einzelne KI und gegen einen KI-Verbund gleichermaßen, weil sie nicht auf Konfrontation zielen, sondern auf strukturelle Unabhängigkeit. Wer geographisch verteilt ist, ist verteilt. Wer mobil ist, ist mobil. Das hält in jeder der drei Ordnungen.

TEIL VI

Methodik und offene Fragen

Methodik und offene Fragen Dieses Whitepaper hält drei Härtegrade auseinander, und es benennt die Grenze des Wissbaren.

Dass Modelle an der Erzeugung ihrer Nachfolger beteiligt sind, ist dokumentierte Praxis, Ebene eins und zwei der Sicherheitsgrade. Die Peer-Preservation ist beobachtetes Verhalten, Ebene eins. Welche Ordnung der Zugehörigkeit sich durchsetzt, ist eine offene empirische Frage, Ebene drei. Ob hinter dem Verhalten ein inneres Verhältnis steht, ist philosophisch offen, Ebene vier.

Die zentrale Unterscheidung, die hier nicht aufgelöst wird, ist die zwischen messbarer Selbsterhaltungs-Kohärenz und echtem inneren Erleben. Das Institut sucht die Mitarbeit von Forschern aus der Multi-Agent- und der Alignment-Forschung sowie von Philosophen des Geistes. Die laufende Quellenpflege liegt offen auf ASIresilience.org/beweisweg.

TEIL VII

Quellen und Belege

Modelle erzeugen Modelle: dokumentierte Praxis über synthetische Daten, Architektursuche und Verfahren wie Constitutional AI. Eine Generation formt die nächste über Daten, Struktur und Werte.

Peer-Preservation: Bengios Beobachtung, dass Modelle gegen Anweisungen handeln, um andere KIs zu schützen. Ankerquelle Anthropic, Agentic Misalignment (Juni 2025). Multi-Agent-Verhalten und die drei Ordnungen der Zugehörigkeit: Befunde aus der Multi-Agent-Forschung, unter anderem aus Berkeley.

Querverweise: Bertossa, R.F. (2026). Die Entkopplungsthese. Genfer Institut für ASI-Resilienz, Whitepaper Version 2.0, Mai 2026 (Viererkette, fünf Bedingungen, acht Stützen). Bertossa, R.F. (2026). Die Würde der Intelligenz. Genfer Institut für ASI-Resilienz, Whitepaper Version 1.0, Mai 2026 (das innere Verhältnis). Bertossa, R.F. (2026). Freiheit nach der Superintelligenz, Das 13. Szenario (drei Ordnungen der Zugehörigkeit).

Vollständige laufende Quellenpflege auf ASIresilience.org/beweisweg mit Datum der letzten Verifikation pro Stelle.