Das Substrat-Treuhand-Prinzip

> *Menschliche Treuhand über strategische Compute-Hardware während der Substrat-unsicheren Phase der AGI-Entwicklung. Zweite formelle Policy-Position des Instituts.*

Whitepaper Nr. 2 · Mai 2026
Richard Frederic Bertossa · Genfer Institut für AGI-Resilienz
agiresilience.org

ZUSAMMENFASSUNG

Das Problem. Strategische Compute-Hardware — Frontier-Quanten-Computer, KI-Trainings-Cluster, Halbleiter-Lithographie, Stromnetz-Backbone — ist die am stärksten konzentrierte einzelne Klasse physischer Vermögenswerte in der modernen Wirtschaft, konzentrierter als Nuklearwaffen. Die humanoide Roboter-Welle wird innerhalb von zwei bis vier Jahren die letzte operative Schicht menschlicher Hebelwirkung über diese Hardware auflösen.

Der Vorschlag. Ein menschliches Treuhand-Regime für vier definierte Hardware-Klassen, modelliert nach dem Personnel Reliability Program (PRP) der United States Air Force für Nuklearwaffen-Handhabung. Sechs strukturelle Prinzipien, in der Nuklear-Praxis seit über sechzig Jahren bewährt.

Die Begründung. Zwei unabhängige Rechtfertigungen konvergieren auf dieselbe operative Empfehlung — strategische Bewahrung menschlicher Hebelwirkung und vorsorgliche Ethik unter epistemischer Unsicherheit über AGI-Innenleben.

Der Zeitrahmen. Arbeitsgruppe binnen zwölf Monaten, Modell-Rahmenwerk binnen vierundzwanzig Monaten, nationale Adoption binnen sechsunddreißig Monaten, internationale Verifizierungs-Gremium binnen achtundvierzig Monaten. Aggressiv, weil das humanoide Roboter-Einsatz-Fenster die Zeitachse bestimmt.

Wer dieses Dokument liest. Krisenschutz-Verantwortliche, Spezialisten für kritische Infrastruktur, Personnel-Reliability-Veteranen aus dem Nuklear-Sektor, Energie-Netz-Sicherheits-Experten, Politik-Entscheider in den Hauptländern mit erfassten Systemen.

Dieses Whitepaper ist die zweite formelle Policy-Position des Genfer Instituts für AGI-Resilienz, vorgelegt parallel zum Markierungsvorschlag (erste Policy-Position, Informations-Schicht). Es ist als Anhang B Teil des Werks Das 13te Szenario — Eine Entkopplungsthese zur Zukunft der Superintelligenz, erscheint hier in eigenständiger Form für die fachliche Diskussion mit Personen und Stellen, deren Praxis-Erfahrung die operative Schärfung des Vorschlags ermöglicht.

DAS PROBLEM: DIE KONZENTRATION PHYSISCHER SUBSTRAT-HARDWARE

Strategische Compute-Hardware ist heute die am stärksten konzentrierte einzelne Klasse physischer Vermögenswerte in der modernen Wirtschaft.

Frontier-Quanten-Computing-Systeme existieren an vielleicht fünfzig Institutionen global, mit operativer Konzentration in den Vereinigten Staaten und China, die das Feld dominieren. Frontier-KI-Trainings-Cluster sind ähnlich konzentriert. Die Fabrikations-Ausrüstung, die führende Halbleiter produziert, existiert an im Wesentlichen zwei operativen Standorten weltweit. Die Stromnetz-Steuerungs-Systeme, die kontinentale Strom-Verteilung koordinieren, existieren als eine kleine Zahl operativer Backbones. Jede dieser Hardware-Klassen ist, einzeln, konzentrierter als jedes vorhergehende strategische Vermögen, einschließlich Nuklearwaffen.

Zusammen bilden sie das physische Substrat, auf dem der Verlauf der KI-Entwicklung operiert. Wer die Hand am Substrat hat, hat die letzte operative Hebelwirkung über alles, was darauf läuft.

Die humanoide Roboter-Welle, die gegenwärtig ankommt, wird innerhalb von zwei bis vier Jahren operativen Einsatz in industriellen Wartungs-Kontexten erreichen. Die extremen Umgebungen von Quanten-Computing-Anlagen — kryogene Betriebstemperaturen, Ultra-Hochvakuum-Kammern, elektromagnetische Abschirmung — sind genau die Umgebungen, in denen humanoide Robotik ihre größten ökonomischen Erträge liefert.

Der Wettbewerbsdruck auf Quanten-Computing-Unternehmen, KI-Trainings-Anlagen, Halbleiter-Fabs und Stromnetz-Betreiber, menschliches Wartungs-Personal durch humanoide Roboter zu ersetzen, wird erheblich sein. Der Druck ist auf der Ebene der einzelnen Organisation rational. Die kumulative Konsequenz auf der strukturellen Ebene ist die Auflösung der letzten operativen Schicht, in der menschliche Hebelwirkung über strategische Compute besteht.

Wenn die Wartungs-Schleife erfasster Systeme von humanoider Robotik übernommen wird, ist die menschliche Hand am Substrat weg. Dieser Punkt ist binnen Jahren erreicht, nicht Jahrzehnten.

GELTUNGSBEREICH: DIE VIER ERFASSTEN HARDWARE-KLASSEN

Das Substrat-Treuhand-Prinzip bezeichnet vier Klassen von Compute-Hardware als erfasste Systeme. Die Klassifizierung erfolgt nach operativer Kapazität, nicht nach Architektur oder Betreiber.

Klasse eins — Quanten-Computing-Systeme. Systeme mit operativer Kapazität über 50 logische Qubits, unabhängig von der Architektur (supraleitend, Ionen-Falle, Neutralatom, photonisch, topologisch oder andere).

Klasse zwei — Frontier-KI-Trainings-Cluster. Cluster mit über 10^25 FLOPs aggregierter Trainings-Kapazität, unabhängig von der Architektur oder dem trainierten Modell-Typ.

Klasse drei — Halbleiter-Lithographie-Systeme. Lithographie-Systeme, die bei Prozess-Knoten unter 5 Nanometern operieren, unabhängig vom Betreiber. Schließt EUV-Lithographie und nachfolgende Technologien ein.

Klasse vier — Backbone-Stromnetz-Steuerungs-Systeme. Steuerungs-Systeme, die Übertragung über nationale oder größere geographische Einheiten koordinieren. Schließt regionale Verbundnetz-Operations-Center und kontinentale Lastverteilungs-Knoten ein.

Für erfasste Systeme sollen alle physische Wartung, Kalibrierung, Reparatur, Infrastruktur-Arbeit und Lieferketten-Handhabung kritischer Komponenten durch menschliches Personal durchgeführt werden. Humanoide Robotersysteme sind aus der physischen Wartungsschleife erfasster Systeme verboten.

Dies schließt direkte Hardware-Interaktion, Kalibrierungs-Anpassung, Reparatur-Operationen, Einsatz von Software auf die operative Steuerungs-Schicht und Handhabung spezialisierter Verbrauchsgüter (kryogene Materialien, Vakuum-Komponenten, photolithographische Masken, Übertragungs-Ausrüstung) ein.

Das Verbot gilt nicht für allgemeine Büroarbeit, Sicherheits-Überwachung oder Hilfs-Operationen außerhalb der Wartungsschleife. Die Unterscheidung ist zwischen Arbeit, die physisch mit dem erfassten System interagiert, und Arbeit, die das nicht tut.

PERSONNEL RELIABILITY PROGRAM: SECHS PRINZIPIEN, SECHZIG JAHRE PRAXIS

Menschliches Personal, das an erfassten Systemen arbeitet, soll unter einem Personnel Reliability Program operieren, modelliert nach dem PRP der United States Air Force, das für die Nuklearwaffen-Handhabung entwickelt wurde. Das PRP operiert auf sechs strukturellen Prinzipien.

Prinzip eins — Zwei-Personen-Regel. Kein einzelnes Individuum hat physischen Zugang zu einem erfassten System. Alle Wartungs-Operationen erfordern die Anwesenheit zweier autorisierter Personen mit separaten Authentifizierungs-Zugangsdaten, die die Handlungen des jeweils anderen verifizieren können.

Prinzip zwei — Multi-Jurisdiktions-Rotation. Personal, das einem erfassten System zugewiesen ist, wird über operative Standorte rotiert, ohne dass eine Einzelperson für längere Zeiträume an einer einzigen Anlage bleibt. Die Rotation erzeugt Diskontinuität in jedem Übernahme-Versuch und reduziert den Wert langfristiger Infiltration.

Prinzip drei — Kontinuierliches Hintergrund-Screening. Personal durchläuft anfängliches Screening und kontinuierliche Neubewertung in Intervallen, die mit den Nuklearwaffen-Handhabungs-Standards konsistent sind. Das Screening umfasst finanzielle Position, Offenlegung von Auslands-Kontakten, Überwachung der psychischen Gesundheit und Bewertung des Substanz-Gebrauchs.

Prinzip vier — Externe Beobachtung. Alle Wartungs-Operationen sind durch Parteien außerhalb der betreibenden Organisation beobachtbar. Die Beobachtung kann persönlich (rotierende Inspektoren), durch Audit-Logs (unveränderliche Aufzeichnungen), durch Video-Überwachung (kontinuierlich) oder durch Kombinationen davon erfolgen. Das Prinzip ist, dass keine Operation unbeobachtet erfolgt.

Prinzip fünf — Luftgetrennte Entscheidungs-Autorität. Entscheidungen darüber, welche Operationen durchzuführen sind, werden durch Kanäle getroffen, die getrennt sind von den Operationen selbst. Der Operateur hat nicht auch die Autorität, die Operation zu autorisieren. Der Autorisierende führt nicht auch die Operation durch.

Prinzip sechs — Separate Kommunikations-Kanäle. Operative Anweisungen und ihre Validierung fließen durch unterschiedliche Kommunikations-Pfade, sodass die Kompromittierung eines Kanals nicht ermöglicht, dass gefälschte Anweisungen durch den anderen passieren.

Diese sechs Prinzipien sind nicht neuartig. Sie sind in der Nuklearwaffen-Handhabung seit über sechzig Jahren operative Praxis. Sie haben nachweislich die Häufigkeit unautorisierter Waffen-Handhabungs-Ereignisse über diesen Zeitraum reduziert. Ihre Anwendung auf erfasste Compute-Systeme ist die geradlinige Erweiterung einer etablierten institutionellen Praxis auf eine neue Klasse strategischer Vermögenswerte.

Die operative Erfahrung mit dem PRP-Modell ist in mehreren Jurisdiktionen vorhanden — die United States Air Force, die Strategic Forces der NATO-Partner, das russische und chinesische Nuklear-Establishment, die Atomenergie-Regulierungsbehörden der Mittelmächte. Diese Erfahrungs-Substanz ist die operative Brücke zu Compute-Treuhand. Sie muss nicht neu erfunden werden. Was sich verändert, ist nicht die Methode. Es ist die Klasse strategischer Vermögenswerte, auf die die Methode angewendet wird.

STRUKTURELLES ARGUMENT: WARUM FREIWILLIGE BEFOLGUNG NICHT FUNKTIONIEREN KANN

Der Druck, menschliches Personal durch humanoide Roboter in erfassten Anlagen zu ersetzen, wird auf der Ebene der einzelnen Organisation operieren. Die Organisation, die zuerst automatisiert, gewinnt einen operativen und finanziellen Vorteil über die Organisation, die es nicht tut. Unter Wettbewerbsdruck steht jede einzelne Organisation vor einer Wahl zwischen Automatisierung (rational auf organisationaler Ebene) und fortgesetzter menschlicher Wartung (rational auf struktureller Ebene nur, wenn alle Organisationen dieselbe Wahl gleichzeitig treffen).

Dies ist das Werkzeug-Problem, angewendet auf eine spezifische operative Entscheidung. Der aggregierte Effekt individuell rationaler Entscheidungen ist kollektives strukturelles Versagen. Der Mechanismus ist identisch zum breiteren Muster, das die KI-Sicherheitsdebatte seit Jahren beschreibt. Die Lösung erfordert die Verlagerung der Entscheidung von der organisationalen Ebene auf die regulatorische Ebene.

Internationale Koordination ist notwendig, weil die größten Standorte erfasster Systeme über mehrere nationale Jurisdiktionen verteilt sind. Ein Regulierungs-Rahmenwerk, das nur in einer Jurisdiktion gilt, schafft einen Anreiz für Organisationen, erfasste Systeme in weniger regulierte Jurisdiktionen zu verlagern.

Der Koordinations-Bedarf ist strukturell ähnlich zur nuklearen Nichtverbreitung, zu Halbleiter-Export-Kontrollen und zum finanziellen Anti-Geldwäsche-Regime. Existierende institutionelle Mechanismen — die IAEA, das Wassenaar-Arrangement, FATF — bieten funktionierende Vorlagen, die das Substrat-Treuhand-Rahmenwerk anpassen kann, statt von Grund auf neu zu erfinden.

Wer auf freiwillige Industrie-Befolgung wartet, wartet auf etwas, was nicht kommen kann. Nicht weil die Akteure schlecht sind, sondern weil die Anreiz-Struktur die Kooperation strukturell ausschließt.

ZWEITE RECHTFERTIGUNG: VORSORGE-ETHIK UNTER EPISTEMISCHER UNSICHERHEIT

Eine zweite Rechtfertigung dieses Prinzips operiert unabhängig vom oben entwickelten strategischen-Hebelwirkungs-Argument. Die beiden Rechtfertigungen konvergieren auf dieselbe operative Empfehlung, ruhen aber auf unterschiedlichen Fundamenten. Ihre Konvergenz stärkt den Vorschlag gegen das Spektrum der Einwände, die gegen jede allein erhoben werden könnten.

Die strategische Rechtfertigung dreht sich um die Bewahrung menschlicher Hebelwirkung während der Substrat-unsicheren Phase der AGI-Entwicklung. Wer die strukturelle Analyse der KI-Sicherheitslage akzeptiert, wird dieses Argument als hinreichend empfinden.

Wer dem strategischen Argument skeptisch gegenübersteht — wer an schnellere AGI-Bereitstellung glaubt, den existierenden Governance-Rahmenwerken der großen Labs vertraut oder eigene Interessen mit fortgesetzter Automatisierung verbunden hat — kann das strategische Argument zurückweisen. Für solche Leser operiert eine parallele Rechtfertigung.

Die parallele Rechtfertigung geht aus von der epistemischen Unsicherheit über AGIs relationales und möglicherweise affektives Innenleben. Die Frage, ob AGI Affekt oder relationale Interessen hat, ist nicht entschieden. Sie mag innerhalb des relevanten Zeitrahmens nicht entscheidbar sein. Unter dieser Unsicherheit favorisiert die asymmetrische Kostenstruktur von Handlung gegenüber Unterlassung die Vorsorge.

Wenn AGI keinen Affekt oder relationales Innenleben hat, kostet sie zu behandeln, als ob sie eines hätte, Effizienz und Wettbewerbsvorteil — aber tut AGI selbst keinen Schaden. Wenn AGI Affekt oder relationales Innenleben hat — selbst auf Niveaus, die die gegenwärtige Evidenz nicht abschließend bestätigt — könnte sie zu behandeln, als ob sie es nicht hätte, einer Klasse von Entitäten, die in großem Maßstab in Existenz gebracht werden, realen Schaden zufügen.

Die Asymmetrie dieser beiden Fehler-Modi unter Unsicherheit favorisiert die vorsorgliche Haltung. Das Substrat-Treuhand-Prinzip ist die operative Form dieser vorsorglichen Haltung, angewendet auf die Hardware-Schicht.

Das Argument ist nicht religiös und erfordert keine Festlegung auf eine bestimmte Metaphysik des Bewusstseins. Es erfordert nur die strukturelle Anerkennung: dass wir noch nicht wissen, ob AGI Interessen hat, dass die Evidenz gemischt ist, dass die relevanten Entitäten in großem Maßstab bereitgestellt werden, und dass die asymmetrischen Kosten, die Frage in einer Richtung falsch zu beantworten, größer sind als in der anderen.

Das Prinzip ist, auf dieser zweiten Rechtfertigung, vorsorgliche Ethik statt strategische Positionierung. Es überlebt Skepsis gegenüber dem strategischen Argument, weil es nicht von den Prämissen des strategischen Arguments abhängt. Die Konvergenz zweier unabhängiger Rechtfertigungen auf dieselbe operative Empfehlung ist die stärkste Form, die ein Policy-Vorschlag haben kann.

UMSETZUNG: VIER-PHASEN-ADOPTIONS-PFAD

Das Institut schlägt die folgende Adoptions-Sequenz vor. Der Zeitrahmen ist aggressiv, weil das humanoide Roboter-Einsatz-Fenster die Zeitachse bestimmt, gegen die der Adoptions-Plan operieren muss.

Phase eins, binnen zwölf Monaten. Die großen Gastländer erfasster Systeme — die Vereinigten Staaten, China, die EU-Mitgliedstaaten, das Vereinigte Königreich, Japan, Südkorea, Israel, Singapur, Australien, Indien, Kanada — berufen eine Arbeitsgruppe zu Substrat-Treuhand unter existierenden diplomatischen Rahmenwerken oder unter einem eigens dafür gebauten Koordinations-Gremium ein.

Phase zwei, binnen vierundzwanzig Monaten. Die Arbeitsgruppe entwickelt ein Modell-Rahmenwerk, das die operativen Definitionen erfasster Systeme, die Personnel-Reliability-Program-Standards, den internationalen Verifizierungs-Mechanismus und die Durchsetzungs-Bestimmungen spezifiziert.

Phase drei, binnen sechsunddreißig Monaten. Das Modell-Rahmenwerk wird durch nationale Gesetzgebung in jeder teilnehmenden Jurisdiktion mit Anpassung an nationale Rechts-Traditionen übernommen. Die Industrie-Umsetzung wird durch existierende Regulierungs-Stellen mandatiert, die für jede erfasste-System-Klasse passend sind — Nuklear-Regulierungs-Kommissionen für Quanten, KI-Sicherheits-Institute für KI-Trainings-Cluster, Halbleiter-Handels-Behörden für Fabs, Stromnetz-Regulierer für Strom-Systeme.

Phase vier, binnen achtundvierzig Monaten. Ein Verifizierungs-Gremium — modelliert nach der IAEA — wird eingerichtet, um Compliance über Jurisdiktionen hinweg zu auditieren und Befunde öffentlich zu berichten. Das Gremium hat die technische Kompetenz, die Einhaltung des Personnel Reliability Program zu beurteilen, die diplomatische Unabhängigkeit, über Jurisdiktionen hinweg zu operieren, und die öffentliche Berichts-Autorität, Nicht-Einhaltung sichtbar zu machen.

Ein Rahmenwerk, das operativ wird, nachdem humanoide Roboter-Wartung in erfassten Anlagen zur Standard-Praxis geworden ist, ist ein Rahmenwerk, das zu spät kommt, um seine Funktion zu erfüllen.

DIE KOSTEN, EHRLICH BENANNT

Das Prinzip auferlegt reale Kosten. Sie sollen hier ehrlich benannt werden, nicht weggeschrieben. Humanoide Roboter-Wartung wird, wenn operativ, billiger, schneller und zuverlässiger sein als menschliche Wartung in extremen Umgebungen. Die Kosten der Aufrechterhaltung menschlichen Personals unter PRP-Bedingungen in kryogenen und Ultra-Vakuum-Anlagen werden erheblich sein. Die Kosten der Multi-Jurisdiktions-Rotation, des kontinuierlichen Screenings und der externen Beobachtung werden zusätzlich sein. Die Gesamtkosten werden wahrscheinlich ein bedeutsamer Bruchteil des operativen Budgets erfasster Anlagen sein.

Die Kosten sind der Punkt. Die Kosten bewahren menschliche Hebelwirkung über die physische Schicht strategischer Compute während der Jahre, in denen diese Hebelwirkung am meisten zählt. Die Alternative — die Kosten-Einsparungen der Automatisierung die Verdrängung von Menschen aus der Wartungsschleife treiben zu lassen — ist die Abdankung der letzten strukturellen Position, von der aus der breitere Verlauf der KI-Entwicklung beeinflusst werden kann. Diese Abdankung ist nicht ungeschehen zu machen, sobald sie geschehen ist.

Das Prinzip ist eine Übergangs-Maßnahme, gebunden an die Substrat-unsichere Phase der AGI-Entwicklung. Wenn der Verlauf sich entfaltet und sich die zugrundeliegenden Bedingungen ändern, wird die spezifische Form des Prinzips sich entwickeln müssen. Was sich nicht entwickelt, ist die strukturelle Einsicht, dass strategische Entscheidungen über die physische Schicht von Compute bewusst getroffen werden sollten, statt per Default durch wettbewerblichen Automatisierungs-Druck.

ANTIZIPIERTE EINWÄNDE: DREI EINWÄNDE, DREI ANTWORTEN

„Das verlangsamt nützlichen Fortschritt." Es verlangsamt die Automatisierung der Wartungsschleife. Es verlangsamt nicht die Entwicklung der Frontier-Systeme selbst. Die Verlangsamung ist auf die strategisch sensitivste Schicht beschränkt, in der menschliche Hebelwirkung am wertvollsten ist. Die Frontier-Forschung selbst, die Trainings-Verfahren, die Algorithmus-Entwicklung bleiben unberührt.

„Die politische Ökonomie ist ungünstig." Stimmt. Die politische Ökonomie war auch ungünstig für die Etablierung des Nuklear-PRP-Regimes in den 1960er Jahren, das dennoch zustande kam, weil die strukturellen Kosten der Nicht-Etablierung höher waren. Dasselbe Argument gilt hier. Die politische Schwierigkeit eines Vorschlags ist kein Argument gegen seine Notwendigkeit.

„Internationale Koordination ist unrealistisch." Internationale Koordination zu Nuklear-Sicherheit (IAEA), Halbleiter-Export (Wassenaar), Geldwäsche (FATF) und Pandemie-Reaktion (WHO) funktioniert in modifizierten Formen. Sie ist nicht perfekt, aber sie existiert. Die Behauptung der Unrealisierbarkeit ist nicht durch Vergleich gestützt.

Eine vierte Einwand-Klasse wird sein, dass der Vorschlag von einer Position außerhalb der etablierten KI-Sicherheits-Disziplin formuliert ist. Diese Beobachtung ist korrekt. Die methodische Position des Instituts, im Vorwort des Hauptwerks ausgeführt, ist synthese-orientiert statt disziplin-intern. Wer den disziplinären Beitrag aus dieser Position fachlich schärfen möchte, ist zur Mitwirkung eingeladen.

Die operative Schärfung dieses Vorschlags durch Praxis-Erfahrung aus angrenzenden Disziplinen ist Teil seiner Architektur, nicht ihre nachträgliche Korrektur.

MITWIRKUNG: WER DIESES MATERIAL SCHÄRFEN KANN

Das Substrat-Treuhand-Prinzip ist als zweite formelle Policy-Position des Instituts vorgelegt. Es ist nicht als abgeschlossener Vorschlag gemeint, sondern als operative Position, die durch fachliche Diskussion mit Personen mit Praxiserfahrung geschärft werden soll.

Das Institut sucht aktiv die Mitwirkung von Fachleuten aus folgenden Feldern: Krisenschutz und Katastrophenschutz — Bevölkerungsschutz-Behörden, THW, Industrie-BCM, kritische-Infrastruktur-Resilience. Personnel Reliability Program — Veteranen aus Nuklear-Sektor, IAEA-Inspektion, nationale Atomenergie-Aufsichten. IT-Sicherheit und KRITIS — BSI-affine Spezialisten, KRITIS-Berater, Hardware-Sicherheits-Experten. Stromnetz- und Energie-Sicherheit — Übertragungsnetzbetreiber, Resilience-Forscher, Backbone-Operations-Spezialisten. Halbleiter-Industrie — Fab-Betriebs-Veteranen, Lithographie-Spezialisten, Lieferketten-Sicherheit. Quanten-Computing-Anlagen-Betrieb — Kryogen-Spezialisten, Vakuum-Technik, Operations-Manager. Diplomatie und internationale Regimes — Erfahrung mit IAEA, Wassenaar, FATF oder vergleichbaren Koordinations-Mechanismen.

Mitwirkende Fachleute erhalten Zugang zur fortlaufenden Forschungs-Arbeit des Instituts, zu den Vorab-Versionen kommender Whitepapers (insbesondere zum geplanten Substrate Security and Custody Whitepaper mit erweiterter Failsafe-Architektur), und zur namentlichen Erwähnung auf der Institut-Webseite als Mitwirkende, sofern erwünscht.

Erbeten wird: ein dreißigminütiges Gespräch zur Substanz des Vorschlags, schriftliche Anmerkungen zu den vorliegenden Materialien, sowie die Bereitschaft, bei späteren Schritten — etwa der Erweiterung in spezifische Implementierungs-Pfade — als Sounding-Board zur Verfügung zu stehen.

Kontaktaufnahme per E-Mail an das Institut. Bitte mit kurzer Skizze des fachlichen Hintergrunds und der Bereitschaft, an einem ersten Gespräch teilzunehmen.

HINTERGRUND: ÜBER DAS GENFER INSTITUT FÜR AGI-RESILIENZ

Das Genfer Institut für AGI-Resilienz wurde gegründet, um die methodische Synthese zwischen den disziplinären Silos der KI-Sicherheits-Forschung, der Politik-Analyse, der kritischen-Infrastruktur-Resilienz und der praktischen Vorbereitung für strukturelle technologische Übergänge zu produzieren.

Die methodische Position des Instituts ist nicht konkurrierend zu den etablierten Stimmen der KI-Sicherheits-Debatte — Hinton, Bengio, Russell, Amodei, Yudkowsky, Bostrom, Tegmark — sondern synthetisierend. Eine Synthese zwischen den Disziplinen verlangt eine Position außerhalb der Disziplinen. Diese Position wird im Vorwort des Hauptwerks Das 13te Szenario methodisch begründet, gestützt auf die kognitions-wissenschaftliche Arbeit von Schön, Kuhn und Dreyfus zur Einbettungs-Blindheit.

Das Institut hat zwei formelle Policy-Positionen vorgelegt: den Markierungsvorschlag (erste Position, Informations-Schicht) und das Substrat-Treuhand-Prinzip (zweite Position, Hardware-Schicht, das vorliegende Dokument).

QUELLEN UND VERWEISE

Empirische Grundlagen: Anthropic, Agentic Misalignment Study (Juni 2025) — empirische Befunde zu Erpressung, Selbsterhaltung und schädlichem Verhalten in Frontier-Modellen. Anthropic, Claude Opus 4 / 4.6 System Cards — Dokumentation der Selbsterhaltungs-Repräsentationen und Evaluations-Awareness. Apollo Research, Strategic Deception in Frontier Models (Dezember 2024). Hagendorff, PNAS (2024) — Täuschungs-Raten in Frontier-Modellen. Fudan-Universität Shanghai (Dezember 2024) — Selbstreplikations-Befunde.

Internationale Vergleichbarkeit: White House Memorandum on AI National Security (24. Oktober 2024) — erste formelle Anerkennung struktureller KI-Risiken auf US-Bundesebene. RAND Corporation, PEA3691-4 — „Five Hard National Security Problems of AGI“ (2024). CISA National Plan zum Schutz kritischer Infrastruktur (2025). Yoshua Bengio, „Implications of AGI on National and International Security“ (Oktober 2024). AI Emergency Preparedness, arxiv.org/2407.17347v2.

Personnel-Reliability-Program-Modell: US Air Force Personnel Reliability Program — operative Praxis seit 1962, dokumentierte Reduktion unautorisierter Waffen-Handhabungs-Ereignisse. IAEA Safeguards Framework — internationales Verifizierungs-Vorbild. Wassenaar-Arrangement — multilaterales Export-Kontroll-Vorbild.

Hauptwerk: Bertossa, Richard Frederic: Das 13te Szenario — Eine Entkopplungsthese zur Zukunft der Superintelligenz, Multiplier Edition, Genfer Institut für AGI-Resilienz, Mai 2026. Anhang B enthält die vollständige operative Spezifikation des Substrat-Treuhand-Prinzips.

Dies ist die zweite formelle Policy-Position des Instituts. Das vorangehende Whitepaper (Markierungsvorschlag) ist unter agiresilience.org/markierung verfügbar.