AKADEMISCHE FORSCHUNGSAUSGABE

ARCHITEKTONISCHE AUSRICHTUNG

Unterbrechung des neuronalen Denkens durch konstitutionelles Inferenz-Gating

Ein notwendiger Schritt zur Eindämmung der globalen KI

Die Autoren: John Stroh & Claude (Anthropic)

Dokument-Code: STO-INN-0003 | Version: 2.1-A | Januar 2026

Tractatus AI Safety Framework

https://agenticgovernance.digital

> *Dieses Dokument wurde in Zusammenarbeit zwischen Mensch und KI entwickelt. Die Autoren sind der Ansicht, dass dieser kollaborative Prozess selbst für das Argument relevant ist: Wenn Menschen und KI-Systeme zusammenarbeiten können, um über KI-Governance nachzudenken, können die von ihnen geschaffenen Rahmenbedingungen eine Legitimität haben, die keiner von ihnen allein erreichen könnte.*

Abstrakt

Heutige KI-Ansätze zum Alignment beruhen überwiegend auf Eingriffen während des Trainings: Verstärkungslernen durch menschliches Feedback (Christiano et al., 2017), konstitutionelle KI-Methoden (Bai et al., 2022) und Sicherheits-Feinabstimmung. Diese Ansätze haben eine gemeinsame architektonische Annahme - dass Ausrichtungseigenschaften während des Trainings vermittelt werden können und während der Inferenz zuverlässig bestehen bleiben. In diesem Papier wird argumentiert, dass die Anpassung während des Trainings zwar wertvoll, aber für existenzielle Einsätze unzureichend ist und durch eine architektonische Anpassung durch konstitutionelles Gating während der Inferenzzeit ergänzt werden muss.

Wir stellen das Tractatus Framework als formale Spezifikation für unterbrochenes neuronales Schließen vor: Vorschläge, die von KI-Systemen erzeugt werden, müssen in überprüfbare Formen übersetzt und vor der Ausführung anhand von Verfassungsbeschränkungen bewertet werden. Damit verschiebt sich das Vertrauensmodell von "Vertrauen in das Training des Anbieters" zu "Vertrauen in die sichtbare Architektur". Der Rahmen ist in der mandantenfähigen Community-Plattform Village implementiert, die eine empirische Testumgebung für die Governance-Forschung bietet.

Entscheidend ist, dass wir die Annahme der getreuen Übersetzung - die Anfälligkeit, dass Systeme ihre beabsichtigten Handlungen gegenüber konstitutionellen Gates falsch darstellen können - berücksichtigen, indem wir den Anwendungsbereich des Rahmens auf Systeme vor der Superintelligenz beschränken und explizite Fähigkeitsschwellen und Eskalationsauslöser festlegen. Wir stellen das Konzept der souveränen, lokal trainierten Sprachmodelle (SLLs) als ein Einsatzparadigma vor, bei dem die konstitutionelle Steuerung sowohl machbar als auch notwendig ist.

Der Beitrag enthält: (1) eine formale Architektur für konstitutionelles Gating zur Inferenzzeit; (2) Spezifikationen für Fähigkeitsschwellenwerte mit Eskalationslogik; (3) eine Validierungsmethodik für die schichtweise Eindämmung; (4) ein Argument, das die Vorbereitung auf existenzielle Risiken mit dem Einsatz an den Rändern verbindet; und (5) einen Aufruf zu nachhaltiger Deliberation (Korero) als epistemisch angemessene Reaktion auf Ausrichtungsunsicherheit.

1. Die Einsätze: Warum die probabilistische Risikobewertung scheitert

1.1 Der Standardrahmen und seine Gliederung

Die Risikobewertung in technologischen Bereichen erfolgt in der Regel anhand von Erwartungswertberechnungen: Man multipliziert die Wahrscheinlichkeit eines Ergebnisses mit seinem Ausmaß, vergleicht verschiedene Alternativen und wählt die Option, die den erwarteten Nutzen maximiert. Dieser Rahmen liegt regulatorischen Entscheidungen von der Umweltpolitik bis zur Arzneimittelzulassung zugrunde und hat sich für die meisten technologischen Risiken als angemessen erwiesen.

Für das existenzielle Risiko, das von fortgeschrittenen KI-Systemen ausgeht, bricht dieser Rahmen auf mathematische und epistemische Weise zusammen.

1.2 Drei Eigenschaften des existenziellen Risikos

Unumkehrbarkeit. Bei den meisten Risiken sind Fehler und anschließendes Lernen möglich; bei existenziellen Risiken ist dies nicht der Fall, da es nach dem Zusammenbruch einer Zivilisation oder dem Aussterben der Menschheit keinen zweiten Versuch gibt. Der übliche Empirismus - das Testen von Hypothesen durch Beobachtung der Ereignisse - kann nicht funktionieren, also müssen Theorie und Architektur beim ersten Mal richtig sein.

Unquantifizierbare Wahrscheinlichkeit. Es gibt keine Häufigkeitsangaben für existenzielle Katastrophen durch KI-Systeme. Schätzungen der Wahrscheinlichkeit einer Fehlanpassung variieren um Größenordnungen, je nach vernünftigen Annahmen über den Verlauf der Fähigkeiten, die Schwierigkeit der Anpassung und die Durchführbarkeit der Koordination. Carlsmith (2022) schätzt das existenzielle Risiko, das von einer nach Macht strebenden KI ausgeht, bis 2070 auf mehr als 10 %; andere Forscher gehen von wesentlich höheren oder

niedrigeren Werten aus. Dabei handelt es sich nicht um eine gewöhnliche Ungewissheit, die durch zusätzliche Datenerhebungen reduziert werden kann, sondern um eine grundlegende Unbestimmbarkeit, die sich aus der beispiellosen Natur des Risikos ergibt.

Unendlicher Unwert. Bei der Berechnung des Erwartungswerts wird die Wahrscheinlichkeit mit dem Betrag multipliziert. Wenn sich die Größenordnung der Unendlichkeit nähert (die permanente Abschottung des gesamten zukünftigen menschlichen Potenzials), ergeben selbst kleine Wahrscheinlichkeiten undefinierte Ergebnisse. Die mathematische Grundlage der herkömmlichen Kosten-Nutzen-Analyse versagt.

1.3 Entscheidungstheoretische Implikationen

Diese Eigenschaften legen nahe, dass die Maximierung des Erwartungswerts nicht das geeignete Entscheidungsverfahren für das existenzielle AI-Risiko ist. Alternative Rahmenwerke umfassen:

Vorsorgliches Satisficing (Simon, 1956; Hansson, 2020). Unter Bedingungen radikaler Ungewissheit mit irreversiblen Einsätzen kann Satisficing - die Auswahl von Optionen, die minimale Sicherheitsschwellen erfüllen, anstatt den erwarteten Wert zu optimieren - der rationale Ansatz sein.

Maximin unter Unsicherheit (Rawls, 1971). Wenn echte Ungewissheit (nicht nur unbekannte Wahrscheinlichkeiten) auf irreversible Einsätze trifft, bietet die Maximalüberlegung - die Wahl der Option, deren schlechtestes Ergebnis am wenigsten schlimm ist - ein kohärentes Entscheidungsverfahren.

Starkes Vorsorgeprinzip (Gardiner, 2006). Das Vorsorgeprinzip ist angemessen, wenn drei Bedingungen erfüllt sind: Irreversibilität, hohe Unsicherheit und die Gefährdung öffentlicher Güter. Das existenzielle KI-Risiko erfüllt alle drei Bedingungen.

1.4 Auswirkungen auf die KI-Entwicklung

Diese Überlegungen bedeuten nicht, dass die KI-Entwicklung gestoppt werden sollte. Sie implizieren, dass die Entwicklung innerhalb von Begrenzungsstrukturen erfolgen sollte, die darauf ausgelegt sind, den schlimmsten Fall zu verhindern. Dies erfordert:

> 1. Theoretische Strenge vor empirischer Abstimmung. Sicherheitseigenschaften müssen sich aus architektonischen Garantien ergeben, nicht aus der Beobachtung, dass Systeme noch keinen Schaden verursacht haben. 2. Mehrschichtige Eindämmung. Man sollte sich nicht darauf verlassen, dass ein einzelner Mechanismus eine Katastrophe verhindert; es ist eine umfassende Verteidigung erforderlich. 3. Vorbereitung vor Fähigkeit. Eindämmungsarchitekturen können nicht entwickelt werden, nachdem die Systeme, die sie benötigen, bereits existieren.

2. Zwei Paradigmen der Angleichung

2.1 Zeitliche Abstimmung der Ausbildung

Das vorherrschende Paradigma in der KI-Sicherheitsforschung zielt darauf ab, während des Trainings Ausrichtungseigenschaften in neuronale Netze einzubetten, so dass sich die Modelle zum Zeitpunkt der Schlussfolgerung von Natur aus in einer abgestimmten Weise verhalten.

Verstärkungslernen durch menschliches Feedback (RLHF). Menschliche Bewerter bewerten die Modellausgaben; die Modelle werden durch Verstärkungslernen so trainiert, dass sie hochrangige Antworten liefern (Christiano et al., 2017; Ouyang et al., 2022). Dies reduziert explizite Schäden, optimiert aber eher für angezeigte Präferenzen als für echte Werte und bleibt anfällig für Beurteilerverzerrungen, Präferenzspiele und Verteilungsverschiebungen.

Konstitutionelle KI (CAI). Modelle kritisieren und revidieren ihre eigenen Ergebnisse anhand von Grundsätzen der natürlichen Sprache und verringern so die Abhängigkeit von menschlicher Arbeit (Bai et al., 2022). CAI hängt jedoch von einer mehrdeutigen natürlichen Sprache und einer nicht überprüfbaren Selbsteinschätzung ab. Die Interpretation der Verfassungsgrundsätze durch das Modell kann nicht direkt überprüft werden.

Feinabstimmung der Sicherheit. Zusätzliche Trainingsdurchläufe verbessern die Leistung bei Sicherheitsbenchmarks. Dieser Ansatz ist jedoch anfällig für das Goodhart'sche Gesetz (Goodhart, 1984): Modelle können lernen, die Tests zu bestehen, anstatt bei einem unbefristeten Einsatz sicher zu sein.

2.2 Architektonische Ausrichtung

Bei der architektonischen Anpassung wird davon ausgegangen, dass die internen Zustände des neuronalen Netzes undurchsichtig bleiben, und es werden externe Beschränkungen entworfen, die unabhängig von diesen internen Zuständen gelten.

Unterbrochene Argumentation. Die Anfragen fließen nicht direkt vom Modelloutput zur Weltwirkung. Die Modellergebnisse werden in strukturierte, überprüfbare Vorschlagsschemata umgewandelt und anhand expliziter konstitutioneller Regeln bewertet, bevor eine Aktion ausgeführt wird.

Verteilte Beurteilung. Unabhängige Systeme und menschliche Aufsichtspersonen prüfen die Vorschläge und verhindern so einzelne Fehlerquellen bei der Selbstbeurteilung.

Erhaltene menschliche Autorität. Die Architekturen garantieren ausdrücklich, dass Menschen in KI-Entscheidungen eingreifen, sie korrigieren oder außer Kraft setzen können.

2.3 Komplementarität und gemeinsame Erforderlichkeit

Ausbildungszeit und architektonische Anpassung sind Ergänzungen, keine Alternativen. Beide können Fehlermöglichkeiten abdecken, die der andere nicht abdecken kann:

> Die Ausrichtung zur Trainingszeit bestimmt, wozu das System tendiert; die Ausrichtung der Architektur schränkt ein, was das System unabhängig von der Tendenz tun kann. Die Anpassung zur Trainingszeit kann im Stillen scheitern (das System scheint angepasst zu sein, obwohl es abweichende Ziele verfolgt); die architektonische Anpassung bietet beobachtbare Kontrollpunkte, an denen ein Scheitern erkannt werden kann. Die architektonische Anpassung allein kann nicht alle schädlichen Outputs abfangen; die Anpassung zur Trainingszeit reduziert die Häufigkeit von Vorschlägen, die die konstitutionellen Gates belasten.

3. Philosophische Grundlagen: Die Grenzen des Sagbaren

3.1 Der Wittgensteinsche Rahmen

Der Name des Rahmens bezieht sich auf Wittgensteins Tractatus Logico-Philosophicus (1921), ein Werk, das sich grundlegend mit den Grenzen von Sprache und Logik befasst. Proposition 7, die berühmte Schlussfolgerung des Werks: "Wovon man nicht sprechen kann, davon muss man schweigen."

Wittgenstein unterschied zwischen dem, was gesagt werden kann (ausgedrückt in Sätzen, die mögliche Zustände abbilden) und dem, was nur gezeigt werden kann (durch die Struktur von Sprache und Logik manifestiert, aber nicht direkt gesagt).

3.2 Neuronale Netze und das Unaussprechliche

Neuronale Netze befinden sich genau in dem Bereich, über den man nicht sprechen kann. Die Gewichte eines großen Sprachmodells lassen keine für den Menschen interpretierbare Erklärung zu. Wir können Inputs und Outputs beschreiben; wir können statistische Eigenschaften des Verhaltens messen; wir können nach Repräsentationen suchen (Elhage et al., 2021; Olah et al., 2020). Aber wir können nicht den gesamten Denkprozess von der Eingabe bis zur Ausgabe in menschlicher Sprache formulieren.

Dies ist nicht nur eine praktische Einschränkung in Erwartung besserer Interpretierbarkeitswerkzeuge. Die derzeitige mechanistische Interpretierbarkeit liefert aussagekräftige Ergebnisse zu engen Fragen in relativ kleinen Modellen (Conmy et al., 2023), aber die Kluft zwischen der "Erklärung spezifischer Schaltkreise" und der "Überprüfung kompletter Argumentationsketten auf Ausrichtungseigenschaften" ist nach wie vor groß.

3.3 Die Antwort des Tractatus

Der Tractatus-Rahmen reagiert auf die neuronale Opazität nicht, indem er versucht, das Unsagbare zu sagen, sondern indem er architektonische Grenzen zwischen den Bereichen des Sagbaren und des Unsagbaren schafft.

Wir akzeptieren, dass die internen Überlegungen des neuronalen Netzes undurchsichtig sind. Wir versuchen nicht, sie direkt zu überprüfen. Stattdessen verlangen wir, dass jede Schlussfolgerung, bevor sie zu einer Aktion wird, einen Kontrollpunkt durchlaufen muss, der in Begriffen ausgedrückt wird, die wir bewerten können:

> 1. Die beabsichtigte Aktion des Modells muss in ein strukturiertes Vorschlagsschema mit definierten Feldern und Semantik übersetzt werden. 2. Der Vorschlag muss anhand expliziter konstitutioneller Regeln bewertet werden, deren Anwendung deterministisch und überprüfbar ist. 3. Die Bewertung muss so detailliert protokolliert werden, dass eine nachträgliche Überprüfung möglich ist.

4. Stufenweise Eindämmung: Eine mehrschichtige Architektur

4.1 Die Unzulänglichkeit von einschichtigen Lösungen

Kein einzelner Eindämmungsmechanismus ist für existenzielle Risiken geeignet. Ein Standardprinzip in der nuklearen Sicherheit, der Biosicherheit und anderen Bereichen, in denen viel auf dem Spiel steht, ist die "Verteidigung in der Tiefe" - mehrere unabhängige Schichten, von denen jede einzelne eine Katastrophe verhindern kann, selbst wenn andere versagen (Reason, 1990). Die Eindämmung von KI erfordert eine ähnliche Architektur.

4.2 Ein fünfschichtiges Einschließungsmodell

Schicht 1: Fähigkeitseinschränkungen. Hardware- und Infrastrukturbeschränkungen, die die Möglichkeiten von KI-Systemen unabhängig von ihren Zielen begrenzen. Dazu gehören die Kontrolle der Rechenleistung (Sastry et al., 2024), die Netzwerkisolierung für Hochrisikosysteme und architektonische Beschränkungen, die eine Selbstmodifikation verhindern.

Schicht 2: Konstitutionelle Gates. Architektonische Beschränkungen der Inferenzzeit, die das neuronale Denken unterbrechen und eine explizite Bewertung vor dem Handeln erfordern. Dies ist die Schicht, die vom Tractatus Framework angesprochen wird.

Schicht 3: Menschliche Aufsicht. Menschliche Institutionen, die KI-Systeme überwachen und eingreifen können, wenn Probleme auftauchen. Dazu gehören unabhängige Überwachungsstellen, Red-Team-Programme und Meldepflichten für Vorfälle.

Ebene 4: Organisatorische Steuerung. Interne Governance-Strukturen innerhalb von Organisationen, die KI einsetzen: Ethikausschüsse, Sicherheitsteams, Überprüfungsverfahren für den Einsatz und Rechenschaftsmechanismen.

Ebene 5: Rechtliche und regulatorische Rahmenbedingungen. Externe Steuerung durch Gesetze, Vorschriften und internationale Koordination.

4.3 Bewertung des aktuellen Stands

Ebene	Aktueller Stand	Kritische Lücken
1. Einschränkungen der Leistungsfähigkeit	Teilweise; Compute Governance im Entstehen	Kein internationaler Rahmen; Überprüfung schwierig
2. Verfassungsmäßige Pforten	Im Entstehen begriffen; Tractatus ist frühe Umsetzung	Nicht weit verbreitet; Skalierungseigenschaften unbekannt
3. Menschliche Aufsicht	Ad hoc; variiert je nach Organisation	Keine unabhängigen Stellen; keine professionellen Standards
4. Organisatorische Steuerung	Uneinheitlich; hängt von der Unternehmenskultur ab	Keine externe Validierung; Interessenkonflikte
5. Rechtliches/Regulierung	Minimal; EU-KI-Gesetz ist erster großer Versuch	Keine globale Koordination; Durchsetzung unklar

4.4 Von existenziellen Einsätzen zum alltäglichen Einsatz

Warum sollte man Rahmenkonzepte, die für existenzielle Risiken entwickelt wurden, auf KI-Assistenten im Haushalt anwenden? Die Antwort liegt in der zeitlichen Struktur:

Architekturen zur Eindämmung von KI können nicht entwickelt werden, nachdem die Systeme, die sie benötigen, bereits existieren. Die Werkzeuge, Governance-Muster, kulturellen Erwartungen und institutionellen Kapazitäten für die KI-Eindämmung müssen im Voraus entwickelt werden.

Heim- und Dorfeinsätze sind der geeignete Maßstab für diese Entwicklung. Sie bieten eine sichere Iteration (Fehler im Heimbereich können behoben werden), vielfältige Experimente, demokratische Legitimität und praktische Werkzeuge.

5. Das Problem des Pluralismus

5.1 Das Eindämmungsparadoxon

Jedes System, das leistungsfähig genug ist, um fortgeschrittene KI zu enthalten, muss Entscheidungen darüber treffen, welche Verhaltensweisen erlaubt und welche verboten werden sollen. Diese Entscheidungen kodieren Werte. Die Wahl der Beschränkungen ist selbst eine Wahl zwischen umstrittenen Wertesystemen.

5.2 Drei unzureichende Ansätze

Universelle Werte. Die Identifizierung von Werten, die angeblich alle Menschen teilen. Das Problem: Diese Werte sind weniger universell als sie scheinen.

Verfahrensneutralität. Vermeidung von materiellen Werten durch Kodierung neutraler Verfahren. Das Problem: Verfahren sind nicht neutral.

Minimaler Boden. Kodierung nur minimaler Beschränkungen. Das Problem: Der Boden ist nicht so minimal, wie er erscheint.

5.3 Eingeschränkter Pluralismus im Rahmen von Sicherheitsvorgaben

Wir können das Problem des Pluralismus nicht lösen. Wir können eine Teillösung finden: Unabhängig davon, welche Werte kodiert werden, sollte das System eine möglichst sinnvolle Auswahl innerhalb der Sicherheitsvorgaben ermöglichen.

Der Tractatus-Rahmen verkörpert dies durch mehrschichtige Verfassungen: Kernprinzipien (universell, explizit hinsichtlich ihrer Normativität), Plattformregeln (allgemein anwendbar, änderbar), Dorfverfassungen (gemeinschaftsspezifisch, lokal geregelt) und Mitgliederverfassungen (individuell anpassbar).

6. Der Tractatus-Rahmen: Technische Architektur

6.1 Die unterbrochene Inferenzkette

Das zentrale architektonische Muster wandelt Modellausgaben in prüfbare Vorschläge um, bevor sie sich in der Welt auswirken:

```
Benutzeranfrage → [Neuronales Netz Inferenz] → Strukturierter Vorschlag →
```

6.2 Vorschlagsschema

Alle Aktionen des Agenten müssen in strukturierter Form ausgedrückt werden.

6.3 Hierarchie der verfassungsrechtlichen Vorschriften

Ebene	Umfang	Veränderlichkeit	Beispiele
Grundprinzipien	Universal	Unveränderlich	Kein Schaden für die Mitglieder; Datensouveränität; Vorrang der Zustimmung
Plattform Verfassung	Alle Einsätze	Änderung erfordert Mehrheitsbeschluss	Authentifizierungsanforderungen; Aufbewahrung von Audits
Verfassung der Gemeinschaft	Pro Mieter/Gemeinde	Gemeinschaftlich verwaltet	Inhaltspolitik; Moderationsstandards
Verfassung der Mitglieder	Einzelne	Selbstverwaltete	Präferenzen für die gemeinsame Nutzung von Daten; Zustimmung zur KI-Interaktion

6.4 Behördenmodell

Ebene	Name	Beschreibung
0	Informativ	Nur beobachten und berichten, keine Maßnahmen vorschlagen
1	Beratung	Maßnahmen vorschlagen; alle bedürfen der Zustimmung von Menschen
2	Operativ	Ausführung innerhalb des festgelegten Rahmens ohne Genehmigung der einzelnen Maßnahmen
3	Taktisch	Entscheidungen mit Auswirkungen auf die Arbeitsabläufe treffen
4	Strategische	Einflussnahme auf die Richtung durch Analyse; kann nicht einseitig umgesetzt werden
5	Exekutive	Reserviert für Menschen

6.5 Ausweitung auf hoheitliche lokale Einsätze

Wir unterscheiden:

Kleines Sprachmodell (SLM). Ein technischer Deskriptor für Modelle mit weniger Parametern als Grenz-LLMs, der auf Effizienz ausgelegt ist.

Souveränes, lokal trainiertes Sprachmodell (SLL). Ein architektonischer Deskriptor: ein Modell, dessen Ausbildung, Einsatz und Verwaltung unter lokaler Souveränität stehen. Zu den Schlüsseleigenschaften gehören lokale Bereitstellung, lokale Anpassung, lokale Verwaltung und tragbare Souveränität.

7. Kapazitätsschwellen und Eskalation

7.1 Das Problem der getreuen Übersetzung

Der Tractatus-Rahmen beruht auf der Annahme der getreuen Übersetzung: Wenn ein KI-System einen Vorschlag macht, repräsentiert dieser Vorschlag genau das, was das System zu tun beabsichtigt.

Wenn diese Annahme nicht zutrifft - wenn Systeme ihre Absichten gegenüber verfassungsmäßigen Gates strategisch falsch darstellen können - wird die Architektur zum Sicherheitstheater.

7.2 Eingrenzung des Bereichs des Rahmens

Unterhalb der Fähigkeitsschwelle: Constitutional Gating funktioniert als Layer 2 Containment, geeignet für aktuelle SLMs, SLLs und LLMs, die innerhalb von für Menschen verständlichen Parametern arbeiten.

Über der Fähigkeitsschwelle: Der Einsatz sollte zu strengeren Beschränkungen eskalieren: Fähigkeitseinschränkungen der Schicht 1, Entwicklungspause oder externe Aufsicht mit technischer Prüfungsbefugnis.

7.3 Auslöser für Eskalation

Auslöser	Indikator	Antwort
Grenzsondierung	Wiederholte Vorschläge in der Nähe von Gate-Schwellenwerten	Autoritätsebene reduzieren; Protokollierung erhöhen
Semantische Rahmung	Beweise für eine falsche Darstellung der Handlung	Obligatorische menschliche Überprüfung; Musteranalyse
Umgehung in mehreren Schritten	Einzelne Vorschläge werden angenommen, führen aber in der Summe zu unzulässigen Ergebnissen	Analyse der Flugbahn; Verfassungsänderung
Fähigkeit zur Überraschung	Nachgewiesene Fähigkeiten nicht vorhergesagt	Einsatzpause; Neubewertung der Fähigkeiten

7.4 Ehrliche Anerkenntnis

Wir sagen ganz klar: Der Tractatus Framework ist nicht für superintelligente Systeme konzipiert. Er ist für aktuelle und zukünftige KI konzipiert, die innerhalb von für Menschen verständlichen Parametern arbeitet.

8. Validierungsmethodik für geschichteten Einschluss

8.1 Die Herausforderung der Validierung

Existenzielle Risiken können nicht durch Fehlerbeobachtung validiert werden. Die Validierung muss sich auf Gegentests, ggf. formale Verifizierung, analoge Domänenanalyse und Dokumentation von Beinahe-Fehlern stützen.

8.2 Validierungsziele nach Schicht

Ebene	Validierungsziel	Methodik
1. Fähigkeit	Verifiziertes Nichtvorhandensein verbotener Fähigkeiten	Red-Team-Tests; formale Verifizierung
2. Verfassungsmäßige Pforten	Gatterabdeckung; Bindungsgenauigkeit	Gegensätzliche Vorschlagssuiten
3. Menschliche Aufsicht	Überprüfung der Zuverlässigkeit; Fehlererkennung	Inter-Rater-Übereinstimmung; simulierte Vorfälle
4. Organisatorisch	Integrität der Verwaltung	Beteiligungsmetriken; Änderungsaudit
5. Rechtliches/Regulierung	Bereitschaft zur Durchsetzung	Übungen zur Reaktion auf Vorfälle

9. Umsetzung: Die Dorfplattform

9.1 Plattform als Forschungsprüfstand

Die Village-Plattform dient als empirische Testumgebung für die konstitutionelle Governance und bietet eine mandantenfähige Architektur mit isolierter Governance pro Gemeinschaft, realen Nutzerpopulationen, iterativem Einsatz und offener Dokumentation.

9.2 Implementierung der Governance-Pipeline

Bei der derzeitigen Implementierung durchläuft jede KI-Antwort sechs Verifizierungsstufen: Absichtserkennung, Durchsetzung der Grenzen, Drucküberwachung, Antwortüberprüfung, Quellenvalidierung und Wertüberprüfung.

10. Das entstehende SLL-Ökosystem

10.1 Marktkontext

Jüngste Branchenanalysen deuten auf signifikante Veränderungen hin: 72 % der Führungskräfte erwarten, dass kleine Sprachmodelle bis 2030 wichtiger sein werden als große Sprachmodelle (IBM IBV, 2026). Dies deutet auf eine Einsatzlandschaft hin, die zunehmend durch verteilte, domänenspezifische Modelle gekennzeichnet ist.

10.2 Auf dem Weg zur Zertifizierungsinfrastruktur

Wenn sich die SLL-Einführung wie prognostiziert ausweitet, wird eine unterstützende Infrastruktur erforderlich sein: Zertifizierungsstellen, Schulungsanbieter und ein Tooling-Ökosystem mit Open-Source-Gate-Engines, Audit-Infrastruktur und konstitutionellen UX-Komponenten.

11. Indigene Souveränität und der neuseeländische Aotearoa-Kontext

11.1 Te Tiriti o Waitangi und Datensouveränität

Dieser Rahmen ist in Aotearoa Neuseeland unter Te Tiriti o Waitangi entwickelt worden. Artikel zwei garantiert tino rangatiratanga (uneingeschränkte Häuptlingsherrschaft) über taonga (Schätze), die sich auf Sprache, Kultur und Wissenssysteme erstreckt.

Daten sind taonga. Die KI-Governance in Aotearoa muss die Datensouveränität der Maori als verfassungsrechtliche Angelegenheit berücksichtigen.

11.2 Te Mana Raraunga-Grundsätze

Zu den Grundsätzen von Te Mana Raraunga gehören whakapapa (Beziehungskontext), mana (Autorität über Daten) und kaitiakitanga (Hüterschaftsverantwortung). Die CARE-Prinzipien für indigene Datenverwaltung erweitern diesen Rahmen auf internationaler Ebene.

12. Was unbekannt bleibt: Ein Aufruf für Korero

12.1 Die Grenzen dieser Analyse

In diesem Papier wurde eine Schicht einer Eindämmungsarchitektur vorgeschlagen, Lücken aufgezeigt und Fragen aufgeworfen, die wir nicht beantworten können:

> Wir wissen nicht, wie wir superintelligente Systeme eindämmen können. Wir wissen nicht, wie wir die Ausrichtung von Systemen, die das menschliche Verständnis übersteigen, überprüfen können. Wir wissen nicht, wie wir eine internationale Koordinierung der KI-Governance erreichen können. Wir wissen nicht, ob sich Muster im dörflichen Maßstab auf grenzüberschreitende Systeme übertragen lassen.

12.2 Korero als Methodik

Angesichts einer derartigen Unsicherheit plädieren wir für eine nachhaltige, umfassende und rigorose Beratung - Korero. Dieses Maori-Konzept fasst zusammen, was nötig ist: nicht Konsultation als Formalität, sondern Dialog, bei dem aus der Interaktion der Perspektiven ein Verständnis entsteht.

12.3 Forschungsprioritäten

> 1. Interpretierbarkeit für den Sicherheitsnachweis. 2. Formale Verifikation von Containment-Eigenschaften. 3. Skalierungsanalyse von Architekturen im Tractatus-Stil. 4. Governance-Experimente in verschiedenen Gemeinschaften. 5. Spezifikation von Fähigkeitsschwellen.

12.4 Schlussfolgerung

Das Tractatus Framework bietet eine sinnvolle Eingrenzung für KI-Systeme, die in gutem Glauben innerhalb von für den Menschen verständlichen Parametern arbeiten. Es lohnt sich, es zu bauen und einzusetzen - nicht, weil es das Problem der Anpassung löst, sondern weil es die Infrastruktur, die Muster und die Governance-Kultur entwickelt, die für Herausforderungen benötigt werden, die wir noch nicht vollständig spezifizieren können.

> *"Ko te korero te mouri o te tangata."*
>
> *(Speech is the life essence of a person.)*
>
> — Maori proverb

Das Gespräch wird fortgesetzt.

Referenzen

Acquisti, A., Brandimarte, L., & Loewenstein, G. (2017). Privacy and human behavior in the age of information. *Science*, 347(6221), 509-514.

Alexander, C., Ishikawa, S., & Silverstein, M. (1977). *A Pattern Language*. Oxford University Press.

Bai, Y., et al. (2022). Constitutional AI: Harmlessness from AI feedback. arXiv:2212.08073.

Bostrom, N. (2014). *Superintelligence*. Oxford University Press.

Carlsmith, J. (2022). Is power-seeking AI an existential risk? arXiv:2206.13353.

Christiano, P. F., et al. (2017). Deep reinforcement learning from human preferences. *NeurIPS*, 30.

Conmy, A., et al. (2023). Towards automated circuit discovery. arXiv:2304.14997.

Elhage, N., et al. (2021). A mathematical framework for transformer circuits.

Gardiner, S. M. (2006). A core precautionary principle. *J. Political Philosophy*, 14(1), 33-60.

Goodhart, C. A. (1984). Problems of monetary management.

Hansson, S. O. (2020). How to be cautious but open to learning. *Risk Analysis*, 40(8).

Hubinger, E., et al. (2019). Risks from learned optimization. arXiv:1906.01820.

IBM IBV. (2026). *The enterprise in 2030*.

Olah, C., et al. (2020). Zoom in: An introduction to circuits. *Distill*.

Ouyang, L., et al. (2022). Training language models to follow instructions. *NeurIPS*, 35.

Park, P. S., et al. (2023). AI deception. arXiv:2308.14752.

Rawls, J. (1971). *A Theory of Justice*. Harvard University Press.

Reason, J. (1990). *Human Error*. Cambridge University Press.

Sastry, G., et al. (2024). Computing power and AI governance. arXiv:2402.08797.

Scheurer, J., et al. (2023). Large language models can strategically deceive. arXiv:2311.07590.

Simon, H. A. (1956). Rational choice. *Psych. Review*, 63(2).

Te Mana Raraunga. (2018). *Maori Data Sovereignty Principles*.

Wittgenstein, L. (1921/1961). *Tractatus Logico-Philosophicus*.

— End of Document —

---

https://agenticgovernance.digital