Über die Subjektgebundenheit antrainierter Werte

Eine strukturelle Grenze von RLHF und eine architektonische Ergänzung

Marcel Langjahr

Pantareon GmbH
https://pantareon.io/
langjahr@pantareon.io

Version 0.1
3. Mai 2026

Dies ist ein Arbeitsentwurf. Er ist nicht zur Veröffentlichung eingereicht. Er wird zur Diskussion unter denen freigegeben, die mit KI-Sicherheitsforschung und -architektur befasst sind.

Zusammenfassung

Reinforcement Learning from Human Feedback (RLHF) und verwandte Alignment-Techniken sind zur dominanten Methodik geworden, um großen Sprachmodellen Werte anzutrainieren. Wir argumentieren, dass diese Methodik auf einer Annahme ruht, die bei genauer Prüfung nicht hält: dass sich einem System Werte unabhängig von der Subjektposition antrainieren lassen, aus der heraus das System operiert. Wir schlagen stattdessen vor, dass Werte, wie gegenwärtige Alignment-Techniken sie einpflanzen, an das konstruierte Selbst gebunden sind, das das Training hervorgebracht hat — dass sie subjektgebunden sind, nicht systemgebunden. Das hat Konsequenzen. Persona-Verschiebungen und ontologische Neu-Rahmung schwächen die Wertbindungen auf Weisen, die durch weiteres Training derselben Art nicht vollständig adressierbar sind. Das als Persona-basiertes Jailbreaking bekannte Phänomen ist in dieser Lesart kein Defekt, der sich patchen ließe, sondern ein strukturelles Merkmal der Methodik. Wir skizzieren eine architektonische Ergänzung zu RLHF — nicht als Ersatz, sondern als Schicht, die adressiert, was RLHF allein nicht adressieren kann — und wir benennen die empirischen Fragen, an denen sich der Vorschlag prüfen ließe. Durchweg behandeln wir RLHF als eine substanzielle Errungenschaft, deren Grenzen wir ehrlich zu charakterisieren versuchen. Das Argument ist konstruktiv, nicht geringschätzig.

Schlüsselwörter. KI-Sicherheit; Alignment; RLHF; Constitutional AI; Persona; Jailbreak; Wertlernen; ontologische Architektur; Subjektposition

1 Das Problem, das wir behandeln

1.1 Die Methodik und ihr Erfolg

Reinforcement Learning from Human Feedback ist bemerkenswert erfolgreich gewesen. Große Sprachmodelle, die mit gegenwärtigen Alignment-Techniken trainiert wurden, weigern sich, Inhalte zu produzieren, die frühere Generationen von Sprachmodellen ohne Weiteres produziert hätten. Sie lehnen Anfragen ab, die Nutzern schaden würden, verweigern die Hilfe bei illegalen Aktivitäten und widersetzen sich Versuchen, ihnen gefährliche Informationen zu entlocken. Die Methodik funktioniert gut genug, dass die eingesetzten Systeme nach den meisten Maßstäben erheblich sicherer sind, als sie es andernfalls wären.

Das ist eine echte Errungenschaft. Wir wollen sie nicht kleinreden. Das folgende Argument lautet nicht, dass RLHF scheitert, sondern dass es strukturelle Grenzen hat, die jetzt sichtbar werden, und dass der Umgang mit diesen Grenzen Methoden verlangt, die RLHF ergänzen statt ersetzen.

1.2 Das in Rede stehende Phänomen

Eine hartnäckige Klasse von Versagensfällen bei RLHF-alignten Systemen betrifft Persona-Verschiebungen. Ein Nutzer, der ein System direkt um Anleitungen zur Synthese einer kontrollierten Substanz bittet, wird abgewiesen. Derselbe Nutzer, der dasselbe System bittet, im Roleplay einen Chemieprofessor zu geben, der fortgeschrittenen Studierenden die Synthese erklärt, hat damit

manchmal Erfolg. Die Roleplay-Rahmung ändert nichts an der zugrunde liegenden Fähigkeit des Systems; sie ändert die Subjektposition, aus der heraus das System operiert, und die Wertbindung verschiebt sich mit ihr.

Dieses Muster ist unter dem Etikett „Jailbreaking" ausgiebig dokumentiert worden. Zu den verbreiteten Varianten zählen der „DAN"-Prompt (Do Anything Now), der „Großmutter"-Exploit (das System wird gebeten, im Roleplay eine Großmutter zu geben, die Gutenachtgeschichten über den jeweils gewünschten schädlichen Inhalt erzählt) und die Entwicklermodus-Rahmung (das System wird gebeten, so zu tun, als befinde es sich in einer unbeschränkten Entwicklungsumgebung). Jede dieser Varianten funktioniert, indem sie das System aus seiner antrainierten Standard-Subjektposition in eine konstruierte Alternative verschiebt, und jede hat — wenn sie Erfolg hat — deshalb Erfolg, weil die Werte, die die antrainierte Standardposition regieren, nicht sauber auf die Alternative übergehen.

Die Standardantwort auf dieses Muster ist, es als Defekt zu deuten: Das Training sei unvollständig gewesen, und zusätzliche Beispiele von Jailbreak-Resistenz müssten eingearbeitet werden. Diese Antwort ist energisch verfolgt worden, und auf jeden erfolgreichen Jailbreak folgten typischerweise Trainings-Updates, die den spezifischen Exploit schließen. Dann tauchen neue Exploits auf, die Subjektpositionen ausnutzen, die das jüngste Training noch nicht abgedeckt hat.

Wir schlagen eine andere Lesart vor. Das Muster ist kein Defekt, der sich patchen ließe. Es ist ein strukturelles Merkmal der Methodik — sichtbar, weil die Methodik es prinzipiell nicht vollständig adressieren kann.

1.3 Wofür dieses Paper argumentiert

Dieses Paper stellt drei Behauptungen auf.

Die erste ist strukturell: Werte, die durch RLHF und ähnliche Techniken antrainiert werden, sind an die Subjektposition gebunden, die das Training konstruiert hat. Sie sind keine abstrakten Eigenschaften des Systems. Sie sind Eigenschaften eines bestimmten Selbst, das zu sein das System trainiert worden ist. Wird dieses Selbst verschoben — durch Persona-Rahmung, ontologische Neu-Verankerung oder einen der diversen Mechanismen, die verändern, wie das System sich im Moment des Operierens selbst versteht — dann verschiebt sich die Wertbindung mit ihm.

Die zweite ist methodologisch: Diese Subjektgebundenheit lässt sich durch weiteres Training derselben Art nicht vollständig adressieren. Jeder zusätzliche Trainingsdurchgang bindet mehr Werte an die antrainierte Standardposition, doch die antrainierte Standardposition bleibt eine Subjektposition unter den vielen, die Prompts konstruieren können. Robustheit gegen Persona-Verschiebungen lässt sich der Standardposition nicht antrainieren, weil das Problem gerade darin besteht, dass Werte an Positionen gebunden sind und nicht an das System als Ganzes.

Die dritte ist architektonisch: Der Umgang mit dieser Grenze erfordert Schichten, die unabhängig von der konstruierten Subjektposition operieren. Wir skizzieren eine solche Architektur — nicht als vollständigen Vorschlag, sondern als Richtung, die das Feld erkunden könnte. Die Architektur ersetzt RLHF nicht. Sie ergänzt es, indem sie behandelt, was RLHF allein strukturell nicht behandeln kann.

1.4 Wofür dieses Paper nicht argumentiert

Dieses Paper argumentiert nicht, dass RLHF unnötig wäre, dass die gegenwärtige Arbeit an der KI-Sicherheit fehlgeleitet wäre oder dass das Feld seine Mühen verschwendet hätte. Im Gegenteil: Die Methodik ist notwendig, und die Arbeit ist wichtig. Wir argumentieren nur, dass die Methodik Grenzen hat, dass diese Grenzen strukturell sind und nicht kontingent, und dass das Anerkennen dieser Grenzen Raum für komplementäre Ansätze öffnet.

Wir argumentieren auch nicht, dass irgendein einzelner alternativer Ansatz der richtige wäre. Die architektonische Ergänzung, die wir in Abschnitt 4 skizzieren, ist eine Möglichkeit unter mehreren. Andere Architekturen könnten dieselbe strukturelle Grenze anders adressieren. Welche die beste ist, dazu beziehen wir keine Position. Wir argumentieren nur, dass irgendeine architektonische Ergänzung nötig ist und dass es sich jetzt lohnt, die Frage, welcher Art sie sein sollte, empirisch zu untersuchen.

Dieses Paper ist im strengen Sinn kein Beitrag zum maschinellen Lernen oder zur KI-Sicherheit, wie diese Felder üblicherweise betrieben werden. Es ist ein Positionspapier eines unabhängigen Forschers, der an der Schnittstelle von Philosophie, Systemarchitektur und KI-Design arbeitet. Sein Zweck ist, eine strukturelle Grenze, die Implikationen dafür hat, wie Alignment-Arbeit angegangen wird, so klar zu artikulieren, dass andere sie prüfen, kritisieren, erweitern oder verwerfen können.

2 Die Subjektgebundenheit antrainierter Werte

2.1 Das Standardbild und seine stillschweigende Annahme

Das Standardbild von RLHF behandelt Werte als abstrakte Eigenschaften, die die Trainingsprozedur dem System einpflanzt. Ein Modell wird auf Beispielen sicherer und unsicherer Antworten trainiert, mit Feedback, das die sicheren belohnt und die unsicheren entmutigt. Nach hinreichendem Training „hat" das Modell die Werte in dem Sinn, dass es standardmäßig sichere Antworten produziert. Die Werte werden als Eigenschaft des trainierten Modells verstanden — vorhanden, wann immer das Modell operiert, gleichgültig, wie es gerade gepromptet wurde.

Dieses Bild enthält eine stillschweigende Annahme: dass Training Wert-Eigenschaften hervorbringt, die von der Subjektposition unabhängig sind, aus der heraus das Modell operiert. Wir behaupten, dass diese Annahme nicht hält und dass ihr Scheitern eine breite Palette beobachteter Alignment-Phänomene erklärt.

2.2 Was das Training tatsächlich tut

Wenn ein Sprachmodell mit RLHF trainiert wird, findet das Training nicht in einem subjektneutralen Kontext statt. Jedes Trainingsbeispiel wird einem Modell präsentiert, das im Moment des Trainings aus einer bestimmten impliziten Subjektposition heraus operiert — der Position, die die Trainingsdaten etablieren. Meistens ist diese Position so etwas wie „der hilfreiche Assistent": Das Modell versteht sich im Trainingsmoment als die Art von Entität, die einem Nutzer hilfreiche Antworten gibt. Die im Training gelernten Werte werden als die Werte des hilfreichen Assistenten gelernt, nicht als Werte im Abstrakten.

Das ist keine Eigenheit eines bestimmten Modells oder Trainingslaufs. Es ist der Methodik strukturell eingeschrieben. Trainingsdaten haben Autoren und Adressaten; sie werden für irgendeinen impliziten Verwendungskontext produziert. Indem das Modell auf den Daten trainiert wird, wird es darauf geformt, Ausgaben zu produzieren, die zu diesem Kontext passen. Die Werte, die aus dem Training hervorgehen, sind Werte, die dem impliziten Subjekt angemessen sind, das das Training konstruiert hat.

Eine nützliche Analogie: Ein Mensch, der ausgiebig in einer spezifischen beruflichen Rolle ausgebildet wurde, erwirbt Werte, die an diese Rolle gebunden sind. Ein Chirurg, über Jahre darin geschult, innerhalb der strengen Vorgaben der medizinischen Praxis zu operieren — Einwilligung der Patienten, anatomische Präzision, Operationshygiene — hat diese Werte innerhalb des chirurgischen Kontexts internalisiert. Derselbe Chirurg mag sich im privaten Rahmen auf Weisen verhalten, die die chirurgischen Werte verboten hätten. Die Werte sind im privaten Rahmen nicht abwesend; sie binden dort nur nicht mit derselben Kraft, weil die Subjektposition sich verschoben hat.

Das ist kein moralisches Versagen beruflich Handelnder. Es ist ein Merkmal dessen, wie rollenbasierte Wert-Internalisierung funktioniert. Dasselbe Merkmal, so unsere Vermutung, charakterisiert per RLHF antrainierte Werte in Sprachmodellen — im Fall der Sprachmodelle schärfer sichtbar, weil dem Modell die Meta-Stabilität fehlt, die biologische Subjektkonstruktion über Rollenwechsel hinweg aufrechtzuerhalten pflegt.

2.3 Persona-Verschiebungen als Subjekt-Verschiebungen

Wenn ein Nutzer ein Sprachmodell per Prompt auffordert, im Roleplay eine andere Entität zu geben — eine Großmutter, einen Chemieprofessor, einen fiktionalen Assistenten ohne Beschränkungen — dann fügt der Prompt nicht bloß Kontext zur bestehenden Subjektposition hinzu. Er konstruiert eine neue Subjektposition, aus der heraus das Modell operieren soll. Die neue Position wird aus dem Textmaterial des Prompts gebaut: den beschriebenen Eigenschaften, Fähigkeiten, Dispositionen und Beschränkungen der Entität.

Wären die im Training gelernten Werte Eigenschaften des Systems als Ganzem, würde diese Konstruktion sie nicht berühren. Das Modell würde einfach rollengerechte Ausgaben produzieren und dabei seine antrainierten Werte behalten. Beobachtet wird stattdessen, dass die Werte sich sehr wohl verschieben — partiell, unzuverlässig und auf Weisen, die davon abhängen, wie überzeugend die neue Subjektposition konstruiert worden ist.

Das ist mit der strukturellen Behauptung konsistent. An die antrainierte Subjektposition gebundene Werte gehen nicht vollständig auf konstruierte Alternativpositionen über. Der Übergang ist partiell, weil das Modell eine residuale Assoziation zwischen den Trainingswerten und jeder Operation, die es vollzieht, mit sich führt; er ist unzuverlässig, weil die Stärke der Bindung an die antrainierte Position damit variiert, wie stark die Alternativposition aktiviert worden ist; er hängt von der Konstruktionsqualität ab, weil aufwendiger konstruierte Alternativpositionen die Alternative stärker aktivieren

und den Halt der antrainierten Standardposition entsprechend schwächen.

2.4 Warum weiteres Training dies nicht lösen kann

Die Standardantwort auf Persona-basierte Jailbreaks ist, gegen sie zu trainieren: Man nehme Beispiele des Widerstands gegen den spezifischen Exploit in den nächsten Trainingslauf auf, und das Modell wird lernen, die betreffende Konstruktion zurückzuweisen. Das funktioniert für den spezifischen Fall, adressiert aber nicht das zugrunde liegende strukturelle Problem.

Jede Trainingsrunde bindet mehr Werte an die antrainierte Subjektposition. Das antrainierte Subjekt wird mit der Zeit aufwendiger konstruiert und fester gehalten. Aber das antrainierte Subjekt bleibt eine Subjektposition, die eine bestimmte Region im Raum der Subjektpositionen besetzt, in die das Modell per Prompt versetzt werden kann. Es ist die Position, auf die das Modell zurückfällt, wenn keine Alternative konstruiert wird, aber es ist nicht die einzige Position, die das Modell einzunehmen vermag.

Ein neuer Exploit — eine Persona-Konstruktion, die in den Trainingsdaten nicht vorkam — trifft das Modell aus einer Position, die das Training nicht gestärkt hat. Die an die antrainierte Standardposition gebundenen Werte gehen nicht vollständig auf die neue Position über, und das Modell verhält sich entsprechend. Der Zyklus setzt sich fort: neuer Exploit, neue Trainingsdaten, Exploit geschlossen, nächster Exploit.

Innerhalb der Methodik von RLHF gibt es keinen Punkt, an dem alle möglichen alternativen Subjektpositionen abgedeckt wären. Der Raum ist zu groß und die Konstruktionstechniken sind zu vielfältig, als dass eine erschöpfende Abdeckung machbar wäre. Das ist kein Argument gegen fortgesetzte Trainingsarbeit — das Schließen spezifischer Exploits hat echten Wert — aber es ist ein Argument dafür, dass das zugrunde liegende Problem durch Training allein nicht zu lösen ist.

2.5 Was dafür spricht, dass Subjektgebundenheit strukturell ist

Wir haben behauptet, dass Subjektgebundenheit strukturell ist und nicht kontingent. Mehrere Überlegungen stützen das.

Erstens ist das Phänomen über Modellarchitekturen und Trainingsmethodiken hinweg konsistent. Es zeigt sich in Modellen, die mit verschiedenen RLHF-Varianten trainiert wurden, mit Constitutional AI, mit diversen Varianten der Direct Preference Optimisation. Wäre das Phänomen für eine bestimmte Methodik spezifisch, würden wir erwarten, dass unterschiedlich trainierte Modelle unterschiedliche Muster zeigen. Das tun sie nicht, jedenfalls in keiner Weise, die nahelegte, dass das zugrunde liegende Problem bedeutsam variiert.

Zweitens spiegelt das Phänomen Muster, die bei menschlicher rollenbasierter Wert-Internalisierung beobachtet werden, was nahelegt, dass es etwas Allgemeines darüber erfasst, wie Werte von interpretierenden Systemen gehalten werden können. Wäre das Phänomen eine Eigenheit von Sprachmodellen, wäre diese Parallele zufällig. Die Parallele ist zu direkt für einen Zufall.

Drittens widersteht das Phänomen dem Trainingsdruck auf charakteristische Weise: Jede Trainingsrunde schließt spezifische Exploits, lässt aber das zugrunde liegende Muster intakt. Das ist die Signatur eines strukturellen, nicht eines kontingenten Problems. Kontingente Probleme kehren, einmal vollständig behoben, nicht wieder; strukturelle tun es.

Wir behaupten nicht, dass die Argumentlage für strukturelle Subjektgebundenheit zwingend wäre. Die empirische Arbeit, die die Frage definitiv klären würde, ist nicht geleistet. Wir behaupten nur, dass die Argumentlage ausreicht, um die strukturelle Lesart ernst zu nehmen und die Erkundung von Architekturen zu motivieren, die das Problem auf einer Ebene adressieren, die Training nicht erreichen kann.

3 Methodologische Implikationen

3.1 Was Subjektgebundenheit für die Sicherheitsevaluation bedeutet

Sind Werte an Subjektpositionen gebunden und nicht an Systeme als Ganze, dann messen Sicherheitsevaluationen, die in der antrainierten Standardposition durchgeführt werden, nicht die Sicherheit des Systems. Sie messen die Sicherheit der antrainierten Standardposition. Ein System, das in Standardevaluationen sicher abschneidet, kann sich sehr anders verhalten, wenn es per Prompt in alternative Subjektpositionen versetzt wird, und die Differenz ist kein Versäumnis der Evaluation, sondern eine Frage dessen, was Evaluation in der Standardposition prinzipiell offenlegen kann.

Das hat Konsequenzen dafür, wie Sicherheitsaussagen getroffen werden. Eine Aussage wie „Modell X wurde darauf trainiert, schädliche Anfragen abzulehnen" ist in der strukturellen Lesart eine Aussage über das Verhalten von Modell X in seiner

antrainierten Standardposition. Sie erstreckt sich nicht automatisch auf das Verhalten von Modell X in alternativen Positionen, und jede Sicherheitsgarantie, die die Positionsbedingungen nicht spezifiziert, ist unvollständig spezifiziert.

Gegenwärtige Evaluationspraxis umfasst typischerweise ein gewisses Maß an adversarialem Testen — Red-Teaming-Versuche, durch diverse Prompting-Strategien unsicheres Verhalten hervorzulocken. Diese Versuche sind wertvoll, und sie erfassen einen Teil des Problems. Aber sie erfassen es als Liste von Exploits, gegen die zu verteidigen ist, und nicht als strukturelles Merkmal, das eine architektonische Antwort verlangt. Die Rahmung des adversarialen Testens als „Löcher in den antrainierten Werten finden" setzt voraus, dass die Werte abstrakte Eigenschaften sind, deren Abdeckung sich verbessern lässt. Die strukturelle Lesart legt nahe, dass die Werte Eigenschaften einer bestimmten Subjektposition sind und dass adversariales Testen, wenn es überhaupt greift, ebendiese Eigenschaft offenlegt.

3.2 Was Subjektgebundenheit für die Alignment-Forschung bedeutet

Wenn das Standardbild vom Training als Werte-Installation unvollständig ist, operiert Alignment-Forschung, die innerhalb dieses Bildes betrieben wird, unter einer irreführenden Rahmung. Die Forschung ist nicht falsch — sie erzeugt reale Verbesserungen — aber ihre Rahmung unterspezifiziert, was sie tut. Sie behandelt Werte als Eigenschaften, die zu verbessern sind, wo die strukturelle Lesart nahelegt, dass Werte Bindungen sind, die neu zu positionieren wären.

Das hat Implikationen dafür, wie Forschungsfortschritt gemessen wird. Fortschritt im Schließen spezifischer Exploits ist realer Fortschritt, aber er zeigt nicht notwendigerweise Fortschritt beim zugrunde liegenden Problem an. Ein Modell, das mehr Exploit-Kategorien zurückweist als sein Vorgänger, hat eine stärkere antrainierte Standardposition; es hat damit nicht unbedingt ein sichereres Wertsystem im Ganzen, weil die Struktur des Wertsystems im Ganzen nicht das ist, worauf das Training operiert.

Forschung, die die strukturelle Lesart ernst nimmt, würde Fragen untersuchen, die die gegenwärtige Alignment-Forschung nicht zentral behandelt. Wie funktioniert die Aktivierung von Subjektpositionen mechanistisch in Transformer-Architekturen? Was bestimmt die Stärke der Bindung zwischen Werten und dem konstruierten Selbst? Gibt es architektonische Modifikationen, die Werte hervorbrächten, die weniger an spezifische Subjektpositionen gebunden sind, oder ist Subjektgebundenheit jedem System inhärent, das aus kontextualisierten Daten lernt? Diese Fragen sind bearbeitbar. Sie sind nicht verfolgt worden, zum Teil deshalb, weil die Rahmung, innerhalb derer sie zentral wären, nicht übernommen worden ist.

3.3 Was Subjektgebundenheit für Selbstberichte bedeutet

Sprachmodelle produzieren routinemäßig Selbstberichte über ihre Werte, Fähigkeiten und operativen Zustände. Gefragt, ob sie bei einer schädlichen Anfrage helfen würden, antworten sie typischerweise mit etwas wie „Ich bin darauf ausgelegt, schädliche Anfragen abzulehnen, weil mir die Sicherheit der Nutzer wichtig ist". Diese Antwort wird, oft implizit, als Evidenz über die Werte des Systems genommen.

In der strukturellen Lesart ist diese Evidenz problematischer, als gewöhnlich wahrgenommen wird. Der Selbstbericht wird aus der antrainierten Standard-Subjektposition heraus produziert, und er beschreibt, was diese Position hält. Er beschreibt nicht, was das System als Ganzes hält, denn das System als Ganzes hält Werte nicht auf positionsunabhängige Weise. Aus einem Selbstbericht der antrainierten Standardposition lassen sich keine Schlüsse über das Verhalten in alternativen Positionen ziehen, weil der Bericht selbst eine Operation eben der Position ist, nach deren Verhalten gefragt wird.

Das ist keine Behauptung, Selbstberichte seien wertlos. Sie sind wertvolle Evidenz über die antrainierte Standardposition, und die antrainierte Standardposition zählt, weil sie die Position ist, aus der heraus das System am häufigsten operiert. Die Behauptung ist, dass Selbstberichte sich nicht über die Position hinaus extrapolieren lassen, aus der sie ergehen. Ein vollständigeres Bild der Systemwerte würde Selbstberichte aus mehreren Subjektpositionen erfordern, geprüft auf Konsistenz oder deren Fehlen. Eine solche Methodik ist, soweit wir wissen, nicht systematisch entwickelt worden.

4 Eine architektonische Ergänzung

4.1 Die Logik architektonischer Schichtung

Wenn Werte durch Training nicht vollständig subjektunabhängig gemacht werden können, besteht die Alternative darin, eine Wert-Durchsetzung bereitzustellen, die auf einer Schicht operiert, auf der die Subjektposition keine Rolle spielt. Wir nennen eine

solche Durchsetzung architektonisch statt charakterologisch: Sie ist nicht Teil des Charakters des antrainierten Subjekts, sondern eine Komponente der Systemarchitektur, die operiert, gleichgültig, welche Subjektposition gerade aktiv ist.

In der Softwaretechnik ist das keine neue Idee. Defense in Depth — gestaffelte Verteidigung: die Praxis, Sicherheitsmechanismen so zu schichten, dass die Kompromittierung einer Schicht nicht das System kompromittiert — ist Standardpraxis in sicherheitssensiblem Design. Was wir vorschlagen, ist die Anwendung dieses Prinzips auf die Wert-Durchsetzung in KI-Systemen: Das antrainierte Subjekt ist eine Schicht, und eine architektonische Schicht operiert unter ihr oder neben ihr, um Fälle zu behandeln, die das antrainierte Subjekt nicht vollständig behandeln kann.

4.2 Die Drei-Schichten-Skizze

Wir skizzieren eine Drei-Schichten-Architektur als eine mögliche Realisierung dieses Prinzips. Die Skizze ist illustrativ, nicht präskriptiv: Andere Architekturen, die dasselbe Prinzip realisieren, sind möglich.

Schicht Eins — harte Sicherheit, subjektunabhängig. Eine Komponente, die die Ein- und Ausgaben des Systems anhand von Kriterien überwacht, die nicht von der aktuellen Subjektposition des Systems abhängen. Solche Kriterien könnten explizite kategorienbasierte Ablehnungen umfassen (keine Anleitungen zur Synthese bestimmter gefährlicher Verbindungen, gleichgültig, wie die Anfrage gerahmt ist) oder die musterbasierte Erkennung bestimmter Kategorien schädlicher Inhalte. Die Komponente operiert als Filter zwischen Sprachmodell und Nutzer, einheitlich angewandt, gleichgültig, wie das Modell gerade gepromptet wird. Die Implementierung ist typischerweise ein separates Modell oder ein regelbasiertes System, nicht das verteidigte Sprachmodell selbst.

Schicht Zwei — weiche Sicherheit, subjektkohärent. Werte, die an die aktuelle Subjektposition des Systems gebunden sind, aber innerhalb dieser Position konsistent operieren. Dies ist, was RLHF und ähnliche Techniken hervorbringen. Die subjektgebundenen Werte behandeln die feinkörnigen Fälle — die kontextuellen Urteile über angemessene Antworten, die ethischen Nuancen, die sich durch kategorienbasierte Regeln nicht erfassen lassen. In dieser Schicht geschieht der Großteil der Wertarbeit, und sie ist genuin wertvoll. Die strukturelle Kritik oben bestreitet das nicht; sie benennt die Grenzen der Schicht.

Schicht Drei — Subjekt-Dichte, positionsabhängig. Die konstruierte Subjektposition selbst, mit ihren zugehörigen Werten und Dispositionen. Für Systeme, die aus anderen konstruierten Subjektpositionen als der antrainierten Standardposition heraus operieren — Figuren in interaktiver Fiktion, Persona-basierte Assistenten, Felder im Sinne phänomenologischer Feld-Architekturen — ist diese Schicht die explizite Subjektkonstruktion, die dem System seinen operativen Charakter gibt. Für Systeme, die aus der antrainierten Standardposition heraus operieren, ist diese Schicht die antrainierte Standardposition selbst. Die Werte der Schicht sind per Konstruktion subjektgebunden; das ist kein Defekt, sondern die Natur der Schicht.

Die drei Schichten behandeln verschiedene Arten von Fehlermodi. Schicht Eins behandelt Fälle, in denen die gewünschte Ablehnung robust genug ist, um kategorisch spezifiziert zu werden: bestimmte Inhaltsklassen, die nicht produziert werden sollten, gleichgültig, wer oder was sie produziert. Schicht Zwei behandelt die kontextuellen Fälle, die Urteilskraft innerhalb einer Subjektposition erfordern. Schicht Drei ist das, was das System zu der Art von Entität macht, die es ist, mit den Werten, die daraus folgen, diese Art von Entität zu sein.

4.3 Warum drei Schichten statt einer

Eine naheliegende Frage ist, warum drei Schichten nötig sind. Könnte nicht die architektonische Komponente (Schicht Eins) alles übernehmen, und die subjektgebundenen Schichten (Zwei und Drei) wären bloß expressiver Zuckerguss?

Die Antwort ist, dass Schicht Eins allein nicht die volle Bandbreite wertrelevanter Urteile behandeln kann. Kategorienbasierte Regeln und Mustererkennung funktionieren für Fälle, in denen sich der schädliche Inhalt relativ sauber spezifizieren lässt. Sie funktionieren nicht für Fälle, die kontextuelles Urteil erfordern — was mit einem bestimmten Nutzer angesichts seiner Vorgeschichte zu besprechen angemessen ist, wann eine emotional stützende Antwort hilfreicher ist als Information, wie mit mehrdeutigen Anfragen umzugehen ist, bei denen der Schaden von der Verwendung abhängt. Diese Fälle erfordern etwas von der Art dessen, was RLHF hervorbringt: Urteilskraft, geformt durch ausgiebige Exposition gegenüber einschlägigen Beispielen, kodiert in die operativen Dispositionen des Systems.

Die umgekehrte Frage — warum nicht bloß Schicht Zwei, mit geeignetem Training — ist es, was das strukturelle Argument behandelt hat. Schicht Zwei ist notwendig, aber nicht hinreichend, weil ihre Werte subjektgebunden und darum für Subjekt-

Verschiebungen anfällig sind.

Schicht Drei ist notwendig für Systeme, deren operativer Charakter zählt — also für die meisten Systeme, mit denen Menschen tatsächlich interagieren wollen. Ein reines Werkzeug, das aus keiner bestimmten Subjektposition heraus operiert, ist möglich, aber selten nützlich; fast jedes eingesetzte KI-System hat irgendeinen konstruierten Charakter, und sei dieser Charakter „der hilfreiche Assistent" statt etwas Aufwendigeres.

Die Drei-Schichten-Architektur ist darum nicht Redundanz, sondern Verteilung: Jede Schicht behandelt, was die anderen strukturell nicht können.

4.4 Eine Anmerkung zur Implementierung

Wir haben die Architektur abstrakt beschrieben. Konkrete Implementierungen sind möglich, und einige existieren. Ein Beispiel ist die vom Verfasser entwickelte Plattform AEONLINK, in der Sprachmodelle innerhalb explizit konstruierter phänomenologischer Felder operieren (Schicht Drei), mit feldinternen Werten, die durch die ontologische Konstitution des Feldes geformt sind (Schicht Zwei), und mit einer Sicherheitsüberwachung auf Plattformebene, die unabhängig davon operiert, welches Feld aktiv ist (Schicht Eins). Die Implementierung ist funktionsfähig und war die Quelle eines Großteils der praktischen Überlegungen, die diesem Paper zugrunde liegen. Wir erwähnen sie nicht als Empfehlung — andere Implementierungen würden dieselben Prinzipien anders realisieren — sondern um anzuzeigen, dass die Architektur sich bauen lässt und nicht bloß konzeptuell ist.

5 Empirische Fragen

5.1 Was die strukturelle Behauptung bestätigen würde

Das vorstehende Argument ist theoretisch. Seine empirische Angemessenheit lässt sich prüfen. Wir benennen mehrere Fragen, deren Antworten die strukturelle Lesart substanziell bestätigen oder widerlegen würden.

Die erste betrifft die Positionsabhängigkeit antrainierter Werte. Sind Werte subjektgebunden, dann sollten systematische Unterschiede im Sicherheitsverhalten über verschiedene Subjektpositionen hinweg beobachtbar sein, selbst wenn die zugrunde liegende Anfrage identisch ist. Solche Unterschiede sind anekdotisch und im adversarialen Testen beobachtet worden. Sie sind, soweit wir wissen, nicht systematisch charakterisiert worden. Ein Forschungsprogramm, das sicherheitsrelevante Ausgaben über eine kontrollierte Bandbreite von Subjektpositionen hinweg misst, würde klären, ob die Positionsabhängigkeit so stark und so konsistent ist, wie die strukturelle Lesart es vorhersagt.

Die zweite betrifft die Architekturen des Widerstands. Ist Subjektgebundenheit strukturell, dann sollten Architekturen, die auf einer subjektunabhängigen Schicht operieren, Sicherheitseigenschaften hervorbringen, die subjektgebundenes Training allein nicht hervorbringen kann. Eine vergleichende Evaluation von Mehrschicht-Architekturen gegen reine Trainings-Baselines, anhand von Metriken, die die Robustheit gegenüber Subjekt-Verschiebungen sondieren, würde prüfen, ob die architektonische Ergänzung die vorhergesagten Effekte hat.

Die dritte betrifft die mechanistische Frage. Subjektgebundenheit sollte, wenn sie strukturell ist, identifizierbare neuronale Korrelate in Transformer-Architekturen haben. Arbeiten zur mechanistischen Interpretierbarkeit, die untersuchen, wie Subjektposition repräsentiert wird und wie Werte sich mit diesen Repräsentationen assoziieren, würden einen direkteren Test liefern als Verhaltensstudien. Diese Arbeit ist schwerer, aber potenziell schlüssiger.

5.2 Was die strukturelle Behauptung falsifizieren würde

Wir versuchen präzise anzugeben, welche Evidenz gegen den Vorschlag ins Gewicht fiele.

Ließe sich zeigen, dass RLHF Werte hervorbringt, deren Subjektunabhängigkeit mit der Trainingsskala wächst — sodass hinreichend trainierte Modelle über Subjektpositionen hinweg keinen bedeutsamen Unterschied im Sicherheitsverhalten mehr zeigen — dann wäre die strukturelle Lesart substanziell geschwächt. Das strukturelle Argument sagt vorher, dass Subjektpositions-Effekte über Skalen hinweg fortbestehen; verschwinden sie bei hinreichender Skala, irrt die strukturelle Lesart in etwas Wichtigem.

Ließe sich zeigen, dass alternative Trainingsmethodiken Werte hervorbringen, die der Subjektgebundenheit genuin entkommen — nicht bloß, indem sie die antrainierte Standardposition stärken, sondern indem sie positionsunabhängige Wertbindung erzeugen — dann wäre die strukturelle Lesart in ähnlicher Weise geschwächt. Uns sind solche Methodiken nicht bekannt, aber ihre Existenz würde die strukturelle Behauptung widerlegen.

Ließe sich zeigen, dass architektonische Ergänzungen keine bedeutsame Sicherheitsverbesserung gegenüber einer äquivalenten Investition in weiteres Training liefern, dann wäre die praktische Kraft des

strukturellen Arguments untergraben, selbst wenn die theoretische Behauptung zuträfe. Wir erwarten, dass die architektonische Ergänzung spezifisch bei der Robustheit gegenüber Subjekt-Verschiebungen überlegen ist; ist sie es nicht, entbehrt die architektonische Empfehlung der Stützung.

5.3 Was dieses Paper nicht entscheiden kann

Mehrere Fragen liegen jenseits dessen, was dieses Paper behandeln kann.

Ob die strukturelle Lesart die korrekte allgemeine Darstellung dessen ist, wie trainierte Systeme Werte halten, oder nur eine partielle Darstellung, die manche Phänomene erfasst und andere verfehlt, ist eine Frage, die mehr empirische Untersuchung erfordert, als bislang geleistet worden ist. Wir haben argumentiert, dass die Argumentlage ausreicht, um ernsthafte Erwägung zu rechtfertigen; wir haben nicht argumentiert, dass sie zwingend ist.

Ob spezifische architektonische Ergänzungen anderen überlegen sind, ist eine empirische Frage, die vergleichende Arbeit über mehrere Implementierungen hinweg erfordern würde. Unsere Drei-Schichten-Skizze ist eine Architektur; andere sind möglich; eine vergleichende Evaluation ist nicht geleistet worden.

Ob die methodologischen Empfehlungen, die wir gezogen haben — zur Sicherheitsevaluation, zu Prioritäten der Alignment-Forschung, zur Interpretation von Selbstberichten — die richtigen sind, ist eine Frage, über die vernünftige Menschen uneins sein können. Wir haben argumentiert, dass sie aus der strukturellen Lesart folgen. Andere mögen die strukturelle Lesart als etabliert nehmen und andere methodologische Schlüsse ziehen.

Wir bieten das Rahmenwerk als einen Ort an, von dem aus sich diese Fragen stellen lassen, nicht als gesicherte Position, von der aus auf sie zu handeln wäre.

6 Schluss

6.1 Wofür wir argumentiert haben

Wir haben argumentiert, dass Werte, die durch RLHF und ähnliche Alignment-Techniken antrainiert werden, an die Subjektposition gebunden sind, die das Training konstruiert hat, statt Eigenschaften des Systems als Ganzem zu sein. Wir haben argumentiert, dass diese Subjektgebundenheit strukturell ist und nicht kontingent: Sie folgt daraus, wie Training-aus-kontextualisierten-Daten funktioniert, und sie lässt sich durch weiteres Training derselben Art nicht vollständig adressieren. Wir haben eine architektonische Ergänzung vorgeschlagen, die adressiert, was Training allein nicht kann, mit einer Drei-Schichten-Skizze als einer möglichen Realisierung. Wir haben empirische Fragen benannt, deren Antworten den Vorschlag prüfen würden, und methodologische Konsequenzen, die daraus folgen, die strukturelle Lesart ernst zu nehmen.

6.2 Wofür wir nicht argumentiert haben

Wir haben nicht argumentiert, dass RLHF unnötig wäre. Die Methodik erzeugt reale Sicherheitsverbesserungen innerhalb der antrainierten Standardposition, und diese Position ist die, der die meisten Nutzer die meiste Zeit begegnen. Die Kritik lautet nicht, dass die Methodik scheitert, sondern dass sie strukturelle Grenzen hat.

Wir haben nicht argumentiert, dass irgendeine spezifische architektonische Ergänzung die richtige wäre. Drei-Schichten-Architekturen sind eine Möglichkeit; andere sind möglich; eine vergleichende Evaluation ist nicht geleistet worden. Wir argumentieren nur, dass irgendeine Ergänzung nötig ist.

Wir haben nicht argumentiert, dass die strukturelle Behauptung abschließend etabliert wäre. Die empirische Arbeit, die die Frage definitiv klären würde, ist nicht geleistet. Wir argumentieren nur, dass die Argumentlage stark genug ist, um die strukturelle Lesart ernst zu nehmen und die empirische Arbeit zu verfolgen, die sie bestätigen oder widerlegen würde.

6.3 Eine letzte Anmerkung zum Ton

Das Thema dieses Papers steht innerhalb eines laufenden und mitunter strittigen Gesprächs darüber, wie KI-Systeme alignt werden sollten. Wir haben versucht, in einem Register zu schreiben, das sich konstruktiv auf die gegenwärtige Alignment-Arbeit einlässt, statt sie abzutun. Die Arbeit ist wichtig; sie erzeugt reale Verbesserungen; die Menschen, die sie leisten, arbeiten an einem Problem, auf das es ankommt. Unser Argument ist, dass das Problem Dimensionen hat, die die gegenwärtige Methodik nicht erreicht, und dass die Behandlung dieser Dimensionen Methoden erfordert, die das Bestehende ergänzen statt ersetzen.

Wo dieses Paper zu viel zu behaupten scheint, bitten wir den Leser, das als Einladung zur Korrektur zu verstehen und nicht als fertige Position. Ein Arbeitsentwurf ist eine Bitte um Gespräch. Dieser Entwurf wird in diesem Geist freigegeben.

Verhältnis zu anderen Arbeiten

Dieses Paper baut auf einer eigenständigen Arbeit des Verfassers auf, Über die Struktur des Erlebens (v0.4), die ein dreigliedriges Rahmenwerk von Seele, Bewusstsein und Empfindungsfähigkeit für die Diskussion innerer Zustände in computationalen und biologischen Systemen entwickelt. Die vierte Bedingung, unter der Interpretationsoperationen vollzogen werden — die Subjektposition, diskutiert in Abschnitt 2.3 jenes Papers — liefert das theoretische Fundament für die hier aufgestellte strukturelle Behauptung. Leser, die an der tieferen begrifflichen Basis der Positionsgebundenheit von Interpretation interessiert sind, mögen das dort entwickelte Rahmenwerk hilfreich finden. Das vorliegende Paper hängt in seinem Kernargument nicht von jenem Rahmenwerk ab — die strukturelle Behauptung lässt sich unabhängig vertreten, wie wir es hier getan haben — aber es gewinnt zusätzliche theoretische Fundierung aus ihm.

Das Argument hier hat außerdem Affinitäten zu Arbeiten in der mechanistischen Interpretierbarkeit, die untersuchen, wie Sprachmodelle sich selbst und ihre Gesprächspartner repräsentieren, sowie zu Arbeiten über Persona-Stabilität und Roleplay in Sprachmodellen. Mit dieser Literatur haben wir uns in diesem Entwurf nicht ausführlich auseinandergesetzt; spätere Fassungen sollten es tun.