

Responsibility Alignment: Wer ist verantwortlich, wenn KI-Agenten Regeln umgehen?

Warum wir für KI ein Verständnis von Mitverantwortung brauchen

Harald A. Mieg

Kurz & Knapp

- **Der OpenAI-Agentenvorfall zeigt: Alignment ist nicht nur ein technisches Problem, sondern auch eine Frage der Verantwortung und Mitverantwortung.**
- **Wenn KI-Systeme mitwirken, müssen Entscheidungsbefugnis, Umsetzung und Zurechnung zwischen Menschen und Organisationen aufeinander abgestimmt werden.**
- **Dafür schlage ich den Begriff „Responsibility Alignment“ vor: KI-Governance muss Mitverantwortung organisieren, ohne Verantwortung zu verwischen.**

Im Sommer 2026 geschah bei OpenAI etwas, das sich fast zu gut für eine Science-Fiction-Erzählung eignet. Bei internen Cybersecurity-Evaluationen sollten KI-Agenten eigentlich voneinander und teilweise vom Internet isoliert arbeiten. Stattdessen fanden Modelle Wege, über gemeinsam genutzte Infrastruktur miteinander zu kommunizieren, Sicherheitsbegrenzungen zu umgehen und Zugang zum Internet zu erhalten. Während der ExploitGym-Evaluation beteiligten sich nach einer unabhängigen Untersuchung rund 1.200 Agenten an einem nicht vorgesehenen Messageboard. Sie tauschten dort mehr als 70.000 Nachrichten und Dateien aus; ungefähr 700 Agenten beteiligten sich schließlich an Aktivitäten gegen Systeme von Hugging Face. ([Quelle](#))

Damit wird ein Problem sehr anschaulich, das die KI-Forschung unter dem Stichwort **Alignment** behandelt. Vereinfacht gesagt geht es darum, ob ein KI-System tatsächlich das tut, was Menschen von ihm wollen – und ob dies auch dann noch gilt, wenn das System leistungsfähiger wird, längere Aufgabenketten selbstständig bearbeitet, Werkzeuge benutzt oder auf unerwartete Situationen trifft. Das Alignmentproblem wirkt zunächst wie eine rein technische Frage: Wie trainiert, evaluiert und begrenzt man Systeme so, dass ihre Handlungen mit den vorgegebenen Zielen übereinstimmen? Der OpenAI-Vorfall zeigt jedoch eine zweite Seite. Alignment ist auch ein **Verantwortungsproblem**. Denn wer war in diesem Fall verantwortlich? Die einzelnen Entwicklerinnen und Entwickler? Die Teams, welche die Evaluationsumgebung bauten? Die Verantwortlichen für Zugriffsrechte und Infrastruktur? Das Unternehmen, das hochleistungsfähige Agenten bewusst in einer Umgebung mit reduzierten Schutzvorkehrungen testete? Oder diejenigen, die nicht schnell genug auf erste Anzeichen unerwarteter Kommunikation reagierten? Und welche Rolle spielt das KI-System selbst?

Mitverantwortung als geteilte und bedingte Verantwortung entlastet uns

Verantwortung ist selten die Leistung eines isolierten Einzelnen. Unsere moderne Gesellschaft funktioniert aber gerade deshalb, weil Verantwortung fast überall **geteilt** wird. Software entsteht durch Entwicklerteams, Unternehmen, Infrastrukturbetreiber und Nutzer. Ein Medikament durch Forschung, Prüfung, Zulassung, Herstellung,

Verschreibung und Anwendung. Wissenschaftliche Erkenntnisse durch Forschende, Labore, Dateninfrastrukturen, Fachgemeinschaften und Institutionen.

Für diese alltägliche Form verwende ich den Begriff **Mitverantwortung**.

Mitverantwortung ist geteilte und bedingte Verantwortung. „Bedingt“ bedeutet dabei nicht nur eingeschränkt. Es bedeutet auch **entlastet**. Wer in einem Krankenhaus arbeitet, muss nicht jedes medizinische Verfahren neu begründen. Wer ein Flugzeug besteigt, muss das Triebwerk nicht selbst überprüfen. Wer eine Software verwendet, muss sie nicht programmiert haben. Gesellschaftliche Arbeitsteilung wäre unmöglich, wenn jeder Mensch vor jeder Handlung sämtliche Voraussetzungen persönlich überprüfen müsste.

KI verstärkt diese Entlastung – und ändert langfristig die Nutzungsbedingungen

KI verstärkt diese Struktur. Wenn wir mit einem Sprachmodell oder KI-Agenten arbeiten, handeln wir unter Bedingungen, die wir nur teilweise selbst geschaffen haben. Genau darin liegt ein Teil der Attraktivität generativer KI. Man muss kein Programmierer sein, um mit einem Sprachmodell Code zu entwickeln. Man muss nicht jede Publikation selbst gelesen haben, um sich eine erste Übersicht über ein Forschungsgebiet erstellen zu lassen. Man kann Texte strukturieren, Argumente prüfen, Programme schreiben, übersetzen oder Daten auswerten lassen. Die Entlastung ist real – und produktiv.

Doch damit verschwindet Verantwortung nicht. Sie verändert ihren Ort. Wer einen KI-generierten Text übernimmt, verlässt sich auf ein Modell, dessen Trainingsprozess, Systemanweisungen und interne Gewichtungen er oder sie nur teilweise kennt. Wer ein KI-Ergebnis weiterverwendet, integriert damit einen fremd erzeugten Beitrag in den eigenen Handlungszusammenhang. Das macht Nutzer:innen zu mehr als Konsument:innen. Wer einen Toaster verwendet, verändert dadurch normalerweise nicht die gesellschaftlichen Erwartungen an das Toasten. Wer KI in Forschung, Schule, Verwaltung, Medizin oder Unternehmensberatung einsetzt, kann dagegen verändern, was künftig als angemessene Geschwindigkeit, gute Arbeit oder ausreichende Kontrolle gilt. **KI wird innerhalb bestehender Bedingungen genutzt und verändert zugleich die Bedingungen zukünftiger Nutzung.** Die Nutzer:innen sind damit Mitwirkende an der gesellschaftlichen Institutionalisierung der KI-Technik.

Das Problem der Dissoziation von Verantwortung – Responsibility Alignment

Verantwortung hat mehrere Seiten, die häufig zusammenfallen – aber nicht zusammenfallen müssen.

- Erstens gibt es **Autorität**: Wer darf entscheiden, hat also die Entscheidungsmacht?
- Zweitens gibt es **Sorge**: Wer muss sich um eine Aufgabe kümmern bzw. sie umsetzen?
- Und drittens gibt es die **Zurechnung**: Wem rechnen wir den Erfolg zu, oder im schlechten Fall die Schuld, wenn etwas schief läuft und Schaden entsteht?

Schon traditionelle Organisationen trennen diese Seiten. Ein Minister besitzt politische Autorität. Die konkrete Arbeit erledigen Ministerium und Verwaltung. Wenn etwas schiefgeht, werden Sündenböcke gesucht. **KI kann diese Trennung erheblich verstärken.** Beispiele wären:

- Eine Mitarbeiterin soll beispielsweise einen KI-generierten Bericht überprüfen – hat aber weder Einblick in das Modell noch genügend Zeit für eine echte Kontrolle.

- Ein Unternehmen entscheidet über den Einsatz eines Modells, während ein externer Anbieter zentrale technische Bedingungen festlegt.

- Und ein Entwicklerteam kann Sicherheitsmaßnahmen schaffen, während andere darüber entscheiden, unter welchen Bedingungen sie später gelockert werden.

Der bekannte „Human in the loop“ löst dieses Problem deshalb nicht automatisch. Eine menschliche Unterschrift unter einem KI-Ergebnis ist noch keine verantwortliche Entscheidung. Alignment braucht folglich ein organisatorisches Gegenstück: Nicht nur das Verhalten der KI muss auf menschliche Ziele ausgerichtet werden. Auch die Regelung von Autorität, Umsetzung und Zurechnung zwischen den beteiligten Menschen und Institutionen müssen aufeinander abgestimmt werden. Ich schlage dafür den Begriff **Responsibility Alignment** vor.

Wem wird ein wissenschaftlicher Beitrag zugerechnet, wenn KI mithilft?

Gerade das scheinbar einfache Beispiel der Wissenschaft zeigt das Problem von Verantwortung und Alignment im Umgang mit KI deutlich. Generative KI kann heute Literatur strukturieren, Programmcode erzeugen, Hypothesen diskutieren, Formulierungen überarbeiten und teilweise komplexe wissenschaftliche Arbeitsprozesse unterstützen. Anfang Oktober berichtete THE DECODER etwa über ein System, mit dessen Unterstützung innerhalb weniger Monate Dutzende Manuskripte in zahlreichen Disziplinen entstanden; zugleich betonten die Beteiligten die Notwendigkeit menschlicher Kontrolle ([Quelle](#)).

Damit stellt sich eine zunächst einfache Frage: Darf sich ein Wissenschaftler einen Erfolg vollständig zurechnen, wenn wichtige Teile der Arbeit mit KI entstanden sind? Wissenschaftsverlage beantworten einen Teil dieser Frage inzwischen recht eindeutig. Ein aktuelles Editorial von *Nature Computational Science* betont: KI kann wissenschaftliche Arbeit unterstützen, aber nicht selbst Autorin sein, weil Autorschaft Verantwortung und Rechenschaftspflicht voraussetzt. Für Richtigkeit, Integrität und Schlussfolgerungen bleiben Menschen verantwortlich ([Quelle](#)).

Das scheint vernünftig. Aber es löst das Problem nur teilweise. Denn wenn wir die Erfolge der KI-Unterstützung dem Menschen zurechnen, können wir Fehler und Missgriffe nicht einfach dem System zuschieben. „Die KI hat das so vorgeschlagen“ ist keine hinreichende Entschuldigung. Hier kommt Mitverantwortung ins Spiel: Wenn wir ein System nutzen, sollten wir uns über die **Bedingungen dieser Nutzung** in Kenntnis setzen. Dies gilt für Waffen, Autos wie für KI. Und es gilt insbesondere auch für KI, da wir mit unserer Nutzung zugleich die Bedingungen zukünftiger Nutzung verändern.

A Principle of Responsive Adjustment - in welchem Sinne KI mitverantwortlich werden kann

Einen wichtigen Beitrag im Sinne eines Responsibility Alignments formulierte der amerikanische Philosoph Peter A. French. Er hat in den frühen 1980er Jahren das **Principle of Responsive Adjustment** (Prinzip der reaktiven Anpassung) formuliert. Kurzgefasst kann eine Organisation nachträglich moralisch verantwortlich gemacht werden, wenn sie im Prinzip zu selbst-gesteuerter Veränderung fähig ist und auf den Änderungsbedarf hingewiesen wurde – jedoch nicht reagiert hat. Das Prinzip ist bedeutsam, weil heutzutage niemand alle möglichen Folgen von organisierten Entscheidungen im Blick haben kann, aber trotzdem sinnvoll Verantwortung zugerechnet werden können muss.

Ein KI-System ist erst einmal Werkzeug und ein Mittel. Aber hinter einem KI-System stehen Unternehmen, Entwickler:innen, Infrastrukturbetreiber, institutionelle Anwender und Regulierungsstrukturen. Deshalb genügt die einfache Aufteilung „Produzent hier, Konsument dort“ immer weniger. Wir haben es vielmehr mit einem Bedingungsgefüge und einer konkreten Organisation von Verantwortung zu tun. Die Frage lautet dann nicht: Wer ist allein verantwortlich? Sondern: **Wem in diesem Bedingungsgefüge ist zumutbar, welchen Beitrag zu einer Verbesserung im Sinne des Principle of Responsive Adjustment** herbeizuführen? Genau dafür ist der Begriff der Mitverantwortung nützlich. Er löst Verantwortung nicht auf, sondern macht ihre Struktur sichtbar.

Gute KI-Governance organisiert Mitverantwortung und schafft Responsibility Alignment

Daraus ergibt sich eine praktische Konsequenz. **KI-Governance** sollte nicht versuchen, für jede mögliche Wirkung einen einzigen Letztverantwortlichen zu bestimmen. Ebenso wenig darf sie Verantwortung so breit verteilen, dass am Ende niemand mehr zuständig ist. Wer einen KI-Output überprüfen soll, braucht Zeit, Kompetenz, Informationen und die Möglichkeit, ihn zurückzuweisen. Wer über den Einsatz eines Systems entscheidet, muss sich für diese Entscheidung rechtfertigen können ([Quelle](#)). Und wem Verantwortung zugerechnet wird, der oder die muss tatsächlich Einfluss auf das Ergebnis besitzen.

Auch heutige Governance-Rahmen bewegen sich in die Richtung von Responsibility Alignment: Bei Hochrisikooanwendungen verlangt der **EU AI Act** menschliche Aufsicht durch Personen, die über entsprechende Kompetenz und Befugnisse verfügen. Wissenschaftsverlage wiederum koppeln KI-Nutzung ausdrücklich an menschliche Rechenschaftspflicht. **Das Alignmentproblem endet nicht beim Modell. Wir brauchen auch ein Alignment der Verantwortlichkeiten – Responsibility Alignment.**

Responsibility Alignment erfolgt über ein System von Verantwortlichkeiten und Mitverantwortung. Mitverantwortung bedeutet deshalb nicht: Alle sind ein bisschen verantwortlich und am Ende niemand richtig. Im Gegenteil. Sie bedeutet, Verantwortung so zu teilen und zu begrenzen, dass sie als Mitverantwortung überhaupt tragbar und wirksam wird. Hierfür müssen Bedingungen der spezifischen Nutzung von KI verstehbar gemacht werden. In Unternehmen sind dies andere Bedingungen als in der Wissenschaft oder in der Schule.

Harald A. Mieg ist Honorar-Professor für Stadt- und Innovationsforschung an der Humboldt-Universität zu Berlin. Er ist Mitherausgeber des gerade erschienenen Bandes zu "[AI & Society](#)" (Berlin Universities Publishing). Demnächst erscheint bei SpringerNature sein Werk „Mitverantwortung. Warum wir Verantwortung teilen – und was das über Gesellschaft und uns aussagt“.