This document was translated from English by machine and may contain mistakes. The English original is at https://www.novaracg.com/2026/04/18/when-control-systems-inherit-the-same-uncertainty-they-are-meant-to-govern/.

Wenn Kontrollsysteme dieselbe Unsicherheit erben, die sie eigentlich beherrschen sollen

Heather Grizzle · April 18, 2026

When AI guardrails are themselves probabilistic models, governance inherits the uncertainty it is meant to control. What real control requires.

KI-Governance wird derzeit anhand dessen beschrieben und vermarktet, was sich am leichtesten visualisieren lässt, und nicht anhand dessen, was am wichtigsten zu analysieren wäre.

Die Diskussion wird von sichtbaren Artefakten beherrscht: Richtlinien-Dashboards, mehrschichtigen Kontrolldiagrammen, beschrifteten Leitplanken und schrittweisen Ausführungsabläufen, die den Eindruck erwecken, eine Anfrage werde vor jeder Aktion sorgfältig geprüft. Diese Darstellungen erwecken den Eindruck von Kontrolle, weil sie traditioneller Systemaufsicht ähneln. Sie verschleiern jedoch die weitaus bedeutsamere Frage, welche Art von Mechanismus innerhalb dieser Schichten die eigentliche Bewertung tatsächlich vornimmt.

In den meisten derzeitigen Architekturen erzeugt ein KI-System einen vorgeschlagenen Handlungsablauf, der dann durch eine Abfolge von Prüfungen geleitet wird, an deren Ende ein weiteres Modell darüber entscheidet, ob die Aktion fortgesetzt werden soll. Diese Bewertungsebene kann je nach Anbieter oder Framework unterschiedlich bezeichnet werden. Sie kann als Policy-Engine, als Risikobewertungsmechanismus oder als regelgestütztes Konfidenz-Ranking-System dargestellt werden. Trotz dieser terminologischen Unterschiede bleibt die zugrunde liegende Struktur konsistent. Ein probabilistisches Modell hat die Aufgabe, die Angemessenheit eines Ergebnisses zu beurteilen, das von einem anderen probabilistischen Modell erzeugt wurde.

Diese Designentscheidung hat erhebliche Auswirkungen, wenn man sie aus der Perspektive von Risiko und Zuverlässigkeit betrachtet. Probabilistische Systeme erzeugen

per Definition unter identischen Bedingungen keine deterministischen, wiederholbaren Ergebnisse. Ihre Ausgaben werden durch statistische Muster geprägt, die aus Trainingsdaten abgeleitet sind, und ihre Bewertungen werden ebenso durch probabilistische Inferenz beeinflusst und nicht durch feste Kriterien. Wird ein solches System zur Steuerung eines anderen Systems desselben Typs eingesetzt, führt die steuernde Funktion nicht zu einer grundlegend anderen Form der Zuverlässigkeit. Stattdessen repliziert sie dasselbe Unsicherheitsprofil auf einer höheren Abstraktionsebene.

Aus Governance-Sicht entsteht dadurch eine zirkuläre Struktur. Das System, das Kontrolle durchsetzen soll, unterliegt derselben Variabilität, derselben Instabilität bei Randfällen und derselben Kontextsensitivität wie das System, das es überwacht. Infolgedessen führt das Vorhandensein zusätzlicher Schichten, Prüfungen oder Wrapper nicht zwangsläufig zu einer sinnvollen Risikominderung. Es kann die Komplexität erhöhen und den Anschein von Sorgfalt erwecken, verändert jedoch nicht das zugrunde liegende Verhalten des Systems in einer Weise, die eine eigenständige Kontrollgrenze schafft.

Dieser Unterschied wird deutlicher, wenn man ihn mit der Funktionsweise von Governance in anderen risikoreichen Bereichen vergleicht. In Finanzsystemen beispielsweise beruhen Freigabeprozesse für Transaktionen auf vordefinierten Regeln, Prüfprotokollen und von Menschen verfassten Kriterien, die nachvollziehbar und rekonstruierbar sind. In sicherheitskritischen technischen Kontexten werden Kontrollsysteme so konzipiert, dass sie andere Zuverlässigkeitseigenschaften aufweisen als die Systeme, die sie überwachen. Der steuernde Mechanismus wird bewusst sowohl in Funktion als auch in Verhalten getrennt, damit er im Falle von Unsicherheit oder Fehlern einen stabilen Bezugspunkt bieten kann.

> Es führt zu Rekursion statt zu Kontrolle.

Aktuelle Muster der KI-Governance stellen diese Trennung häufig nicht her. Stattdessen legen sie probabilistische Beurteilung über probabilistische Generierung und behandeln die resultierende Struktur als Kontrollsystem. Dieser Ansatz geht davon aus, dass das Hinzufügen weiterer Bewertungsschritte innerhalb derselben Zuverlässigkeitsklasse zu Governance-Ergebnissen führt. In der Praxis führt dies jedoch zu Rekursion statt zu Kontrolle, da jede Schicht auf demselben Typ von Denkprozess beruht und denselben Arten von Fehlern ausgesetzt ist.

Die Konsequenz ist nicht rein theoretisch. Wenn ein System ein Ergebnis hervorbringt, das Schaden verursacht, lautet die darauffolgende Frage nicht, ob an der Entscheidung mehrere Schichten beteiligt waren. Die Frage ist, ob irgendeine dieser Schichten als echte Kontrolle fungierte, das heißt, ob sie mit einem anderen Grad an Determinismus, Nachvollziehbarkeit und Verantwortlichkeit arbeitete als das von ihr gesteuerte System. Wenn alle Schichten dieselbe probabilistische Grundlage teilen, fehlt dem System ein klarer Punkt, an dem Verantwortung verankert oder Entscheidungen zuverlässig rekonstruiert werden können.

> **DER ENTSCHEIDENDE TEST**
>
> Governance definiert sich durch die Einführung eines Mechanismus, der mit einem anderen Zuverlässigkeitsprofil arbeitet als das System, das gesteuert wird.

Die sich aus dieser Analyse ergebende Position ist eindeutig. Governance wird nicht durch die Anzahl der in einem System vorhandenen Kontrollen definiert, noch durch die visuelle Komplexität seiner Architektur. Governance definiert sich durch die Einführung eines Mechanismus, der mit einem anderen Zuverlässigkeitsprofil arbeitet als das System, das gesteuert wird. Ohne diese Unterscheidung stellen zusätzliche Schichten keine Kontrolle dar. Sie dehnen lediglich dieselbe Unsicherheit auf mehr Komponenten aus.

Wenn eine Kontrollschicht sich genauso verhält wie das System, das sie eigentlich regulieren soll, erreicht die Architektur keine Governance im eigentlichen Sinne. Es entsteht eine rekursive Struktur, in der Unsicherheit Unsicherheit bewertet und die Grenze zwischen Handlung und Aufsicht zusammenbricht. Genau in diesem Zusammenbruch liegt das eigentliche Risiko, unabhängig davon, wie strukturiert das System an der Oberfläche erscheinen mag.

TIPS

Know something about a sign language AI system that the public should know? Send it to tips@novaracg.com. Every tip is read by me and kept confidential. If you work in the UK, writing to us is not by itself a protected disclosure under the Public Interest Disclosure Act 1998, so before you share employer information, check your options with Protect (protect-advice.org.uk), the free whistleblowing advice line.

Zitierweise

Grizzle, H. M. (2026, April 18). Wenn Kontrollsysteme dieselbe Unsicherheit erben, die sie eigentlich beherrschen sollen. Novara Consulting Group. https://www.novaracg.com/de/2026/04/18/when-control-systems-inherit-the-same-uncertainty-they-are-meant-to-govern/