

Organisationsbionik für KI-Sicherheit

Arbeitsfassung 0.2

19. September 2026

1. Ausgangspunkt

Ich vermute, dass wir bei KI mit einem falschen Grundbild arbeiten.

Wir sprechen davon, dass OpenAI, Anthropic, Google und andere KI-Modelle „bauen“. Das legt nahe, dass wir es mit etwas zu tun haben, das im Wesentlichen konstruiert wird und der Logik von Technik folgt: Man fügt Komponenten hinzu, verbessert einzelne Bauteile und versucht anschließend, das Gesamtsystem sicher zu machen.

Ich glaube, dass dieses Bild irreführend, mindestens aber unvollständig ist.

Eine komplizierte Maschine kann gebaut werden. Man verwendet Schrauben, Normteile und Komponenten mit bekannten Eigenschaften. Die Eigenschaften der fertigen Maschine sind mit sehr großer Sicherheit aus den verwendeten Komponenten vorhersehbar. Falls dies einmal anders ist und z.B. eine Maschine stärker vibriert als angenommen, so stellt dies ein überschaubares Problem dar. Bei modernen KI-Systemen werden Eigenschaften nicht einfach wie Bauteile eingesetzt. Modelle werden trainiert, selektiert, verworfen, weiterentwickelt, nachtrainiert, evaluiert und auf bestimmte Eigenschaften hin optimiert. Man weiß nicht in jedem Fall exakt, welche Kombination von Eigenschaften am Ende entsteht. Ich möchte dazu einen saloppen (etwas überzeichneten) Vergleich verwenden: Es macht für das Weltbild eines Jugendlichen einen Unterschied, in welcher Reihenfolge er Bücher liest: Liest er zuerst Hayeks „Weg zur Knechtschaft“ und danach „Das Kapital“, so kann dies eine andere Prägung bewirken, als wenn er zuerst Marx liest und danach Hayek.

Aus diesem Blickwinkel haben wir es nicht mit „bauen“ zu tun, sondern mit der gezielten Kombination von komplexen Komponenten mit teilweise emergenten Eigenschaften - in diesem Sinn erscheint mir „züchten“ als das präzisere und deutlich besser passende Bild.

Das ist nicht bloß eine sprachliche Verschiebung. Wenn man KI als etwas betrachtet, das gezüchtet wird, stellen sich völlig andere Sicherheitsfragen.

2. Das eigentliche Problem

Es gibt ein KI-Wettrennen.

Die beteiligten Unternehmen und Staaten haben starke Anreize, leistungsfähigere Systeme schneller zu entwickeln als ihre Wettbewerber.

Damit entsteht ein strukturelles Problem:

Sicherheitsbedenken werden zwar von unterschiedlichen Akteuren regelmäßig geäußert, da sie aber aus der Perspektive von „KI bauen“ zwangsläufig bedeuten, langsamer zu werden, werden sie tendenziell eher nicht zum Kern der Überlegungen gemacht.

Ebenso erscheint mir eine Sicherheitsarchitektur, die nur funktioniert, wenn alle Akteure gleichzeitig auf Geschwindigkeit verzichten, als interessantes Gedankenexperiment, dass aber wenig robust ist und mit der Natur des Menschen kaum vereinbar ist.

Das eigentliche Problem lautet daher:

Wie lassen sich Geschwindigkeit, Sicherheit und die zunehmende Macht der Systeme miteinander vereinbaren?

Momentan sehe ich dafür keine überzeugenden Konzepte.

Die naheliegende Reaktion lautet im wesentlichen: stärkere Modelle bauen und anschließend mehr Alignment, mehr Tests und mehr Kontrolle hinzufügen.

Wenn man an den Punkt anlangt, an dem man sich die Frage stellt:

Wie kontrollieren wir ein System mit sehr weitreichenden Fähigkeiten?

Ist man schon vor langer Zeit an der Weggabelung:

Welche Fähigkeiten sollten überhaupt im selben System vorhanden sein?

vorbeigelaufen.

3. Organisationsbionischer Zugang

Ich befasse mich seit über 20 Jahren mit Organisationsbionik und möchte einen Gedankenweg beisteuern.

Der verwendete bionische Prozess:

1. Problem definieren.
2. Nach einem System suchen, das ein strukturell ähnliches Problem bereits lösen musste.
3. Dieses System analysieren.
4. Die dort gefundenen Mechanismen abstrahieren.
5. Prüfen, was davon auf das Ausgangsproblem übertragen werden kann.

Das Ausgangsproblem ist die Sicherheit während des KI-Wettrennens.

Gesucht wird daher ein System, in dem leistungsfähige Eigenschaften stark optimiert werden, gleichzeitig aber die Kombination dieser Eigenschaften selbst zum Problem werden kann.

Die passende biologische Analogie habe ich in der Zucht von Tieren (insbesondere in der Zucht von Gebrauchshunden) gefunden.

Der Mensch hat seit sehr langer Zeit Erfahrung damit, Eigenschaften gezielt zu verstärken, unerwünschte Nebenwirkungen zu beobachten und bestimmte Merkmale miteinander zu kombinieren oder gerade nicht miteinander zu kombinieren.

3.1 Exkurs über Hundezucht

Hunde wurden nicht zum „Superhund“ hin gezüchtet, sondern Züchter hatten immer Eigenschaftenbündel vor Augen, die den Hund für eine bestimmte Aufgabe besonders geeignet, bzw. wirksam machten, gleichzeitig aber dafür sorgten, dass der Hund kontrollierbar blieb. Wir assoziieren große Hunde – wie Bernhardiner - mit Gemütsruhe dies ist aber keine(!) Folge der Größe des Hundes. Große Hunde sind nicht automatisch ruhig – es wäre durchaus möglich (aber keinesfalls wünschenswert) einen Hund von der Größe eines Bernhardiners zu züchten – der das Gemüt eines Chihuahua hätte.

4. Warum Zucht?

Bei der Zucht stellt sich daher nicht nur die Frage:

Welche Eigenschaft will ich verstärken?

Sondern auch:

Welche anderen Eigenschaften kommen damit mit?

Welche Kombinationen funktionieren?

Welche Kombinationen erzeugen unerwünschte Resultate?

Welche Eigenschaften hält man besser in getrennten Linien?

Genau diese Fragen scheinen mir für KI interessant zu sein.

Die Vorstellung, dass man eine Reihe jeweils nützlicher Eigenschaften einfach addiert und damit ein immer besseres Gesamtsystem erhält, erscheint mir sehr von technischem Denken geprägt und ist mir verdächtig wenn man mit Systemen zu tun hat, die bereits Emergenz zeigen.

Eine Eigenschaft kann für sich (sehr) erwünscht sein und in Kombination mit einer anderen Eigenschaft ein Problem erzeugen.

Das wäre ein klassischer Zuchtgedanke.

5. Jagdhunde als stärkeres Beispiel

Betrachten wir historischen Jagdhunde genauer:

Dort gab es funktional sehr unterschiedliche Typen.

Ein Lymer (oder Limer) war auf das Finden und Verfolgen einer Spur spezialisiert. Es waren sehr aufgeweckte, neugierige Hunde.

Ein Alaunt war ein schwerer Greifer. Er hatte beeindruckende Muskel und ein schweres Gebiss

Andere Hunde waren für Geschwindigkeit oder Verfolgung zuständig.

Das Jagdsystem konnte daher insgesamt sehr leistungsfähig sein, ohne dass jeder Hund sämtliche Eigenschaften besaß.

Vereinfacht:

finden → verfolgen → stellen → greifen

Die Funktionen waren verteilt.

Das ist für mich der relevante Punkt.

Ein Lymer mit sehr guter Fährtenarbeit ist eine Sache.

Ein Alaunt mit großer physischer Wirkung ist eine andere.

Die Kombination beider Eigenschaftsbündel in einem Tier erscheint mir deutlich problematisch:

Ein Hund, der

- selbstständig ein Ziel finden kann
- die Spur hartnäckig verfolgt
- das Ziel erreicht
- am Ziel große physische Wirkung entfaltet

bedeutet ein völlig andere Risikoprofil, als zwei Hunde auf die sich diese Eigenschaften verteilen.

6. Herdenschutz

Beim Herdenschutz wird die Sache noch deutlicher.

Hütehund und Herdenschutzhund haben teilweise gegensätzliche Aufgaben.

Der Hütehund soll stark auf Bewegungen der Herde reagieren. Er soll schnell, aufmerksam und eng steuerbar sein.

Der Herdenschutzhund soll gerade nicht auf jede Bewegung innerhalb der Herde reagieren. Er muss unterscheiden, was zum normalen Herdenverhalten gehört und was eine äußere Gefahr darstellt. Gleichzeitig braucht er mehr Selbständigkeit und körperliche Wirkung.

Das Interessante daran ist:

Hier werden dabei aber nicht nur die verschiedene Aufgaben verteilt. Es werden sogar die Reaktionslogiken getrennt. Der Hütehund interagiert mit den Schafen. Der Herdenschutzhund ignoriert die Schafe und hat seinen Fokus nach Aussen gerichtet.

Was beim einen Hund erwünscht ist, wäre beim anderen problematisch.

Die Lösung besteht daher gerade nicht darin, einen Hund zu züchten, der alle gewünschten Eigenschaften maximal vereinigt.

Man bewahrt bestimmte Eigenschaften in getrennten funktionalen Linien auf.

7. Der zentrale Gedanke

Daraus ergibt sich für mich die eigentliche Übertragung:

Nicht jede erwünschte Eigenschaft sollte mit jeder anderen erwünschten Eigenschaft kombiniert werden.

Bei KI scheint die Entwicklung dagegen in Richtung immer stärkerer Fähigkeitsbündel zu gehen.

Modelle sollen gleichzeitig:

- besser programmieren,
- besser planen,
- länger autonom arbeiten,
- bessere Werkzeuge verwenden,
- überzeugender kommunizieren,
- komplexere soziale Situationen verstehen,
- mehr Systeme bedienen,
- Informationen selbständig beschaffen,
- Ziele über mehrere Schritte verfolgen.

Jede dieser Fähigkeiten ist für sich sinnvoll und erleichtert die Arbeit – wie die Eigenschaften von Gebrauchshunden.

Aber daraus folgt nicht, dass ihre großzügige Kombination ebenfalls sinnvoll ist.

Die Sicherheitsfrage müsste deshalb lauten:

Welche Kombinationen wollen wir überhaupt?

Oder vorsichtiger:

Welche Kombinationen brauchen wir überhaupt?

8. KI-Risiko ist wahrscheinlich kombinatorisch

Ich vermute, dass ein Teil des KI-Risikos nicht additiv, sondern kombinatorisch ist.

Programmierungsfähigkeit ist eine Fähigkeit.

Täuschungsfähigkeit ist eine andere.

Tool-Zugriff ist eine weitere.

Langfristige Planung ebenfalls.

Autonome Zielverfolgung wieder eine andere.

Das Risikoprofil eines Systems mit

Programmierungsfähigkeit + Täuschungsfähigkeit + Tool-Zugriff + langfristiger Planung + autonomer Zielverfolgung

ist möglicherweise qualitativ völlig anders als das Risikoprofil der einzelnen Fähigkeiten.

Damit wäre die relevante Einheit der Sicherheitsanalyse nicht nur die Capability, sondern die **Fähigkeiten-Kombination**.

9. Die übliche Entwicklungslogik könnte falsch herum sein

Momentan scheint mir die Entwicklungslogik ungefähr so zu funktionieren:

1. möglichst viele leistungsfähige Eigenschaften erzeugen,
2. diese Eigenschaften in einem möglichst universellen Modell bündeln,
3. anschließend versuchen(!), dieses Gesamtsystem sicher zu machen.

Die Zuchtanalogie legt einen anderen Weg nahe:

1. Eigenschaften gezielt entwickeln,
2. beobachten, welche weiteren Eigenschaften damit gekoppelt sind,
3. Kombinationen prüfen,
4. bestimmte Eigenschaften getrennt halten,
5. nur jene Kombinationen zulassen, die tatsächlich gebraucht werden.
6. Optional: das Modell erziehen (und nicht nur trainieren).

Damit würde Sicherheit bereits bei der Architektur der Fähigkeiten beginnen und nicht erst beim Verhalten des fertigen Systems.

10. „Silos“ bzw. Zuchtlinien

Mein vorläufiges Bild dafür sind getrennte Silos beziehungsweise „Zuchtlinien“.

Bestimmte Fähigkeiten werden sehr weit entwickelt, aber nicht automatisch mit allen anderen Fähigkeiten verbunden.

Ein System kann beispielsweise hervorragend programmieren, ohne die Fähigkeit dem User zu schmeicheln.

Ein anderes System kann Zwischenschritte zu einem vorgegebenen Ziel planen, ohne selbst die Ausführungskompetenz zu besitzen.

Ein weiteres kann externe Aktionen ausführen, aber nur innerhalb eines eng begrenzten Rahmens und ohne eigene langfristige Zielbildung.

Das Gesamtsystem kann dadurch sehr leistungsfähig bleiben – und wahrscheinlich durch die Spezialisierung sogar noch an Leistungsfähigkeit gewinnen.

Aber die einzelnen Komponenten besitzen nicht zwangsläufig die vollständige Wirkungskette.

11. Die abstrahierte Wirkungskette

Die Trennung könnte entlang folgender Kette erfolgen:

wahrnehmen → Ziel auswählen → planen → ausführen → Wirkung erzeugen

Ich sehe keinen zwingenden Grund, warum ein einzelnes autonomes System sämtliche Stufen kontrollieren muss.

Genau hier ist die bionische Übertragung aus der Jagd besonders passend:

Der Suchhund findet.

Der Verfolger verfolgt.

Der Greifer greift.

Der Mensch hält die übergeordnete Kontrolle – und kann jederzeit die Prozesskette unterbrechen, bzw. am weiteren Ablauf hindern.

Nur durch das Gesamtsystem entsteht die vollständige Wirkungskette – sie befindet sich nicht im einzelnen Hund.

12. Das ist keine Forderung nach schwächerer KI

Ich glaube nicht, dass ich in diese Richtung falsch verstanden werde, aber dieser Punkt ist für das Ausgangsproblem des KI-Wettrennens wesentlich. Die Idee lautet nicht:

Wir sollten weniger leistungsfähige KI entwickeln.

Das ist unrealistisch und unter Wettbewerbsbedingungen ohnedies keine Lösung.

Die Idee lautet:

Wir können Fähigkeiten sehr weit entwickeln, ohne alle Fähigkeiten im selben System zu vereinigen.

Damit könnten verschiedene Linien parallel weiterentwickelt werden.

Programmiermodelle können besser werden.

Planungsmodelle können besser werden.

Recherchemodelle können besser werden.

Andere Spezialisten ebenfalls.

Die Sicherheitsfrage liegt dann stärker in der Kombination.

Damit könnte sich der Zielkonflikt zwischen Geschwindigkeit und Sicherheit zumindest teilweise verschieben.

13. Gezielte Zucht statt wahlloser Akkumulation

Die organisationsbionische Lösung wäre daher nicht, KI „weniger mächtig“ zu machen.

Sie wäre, Eigenschaften gezielter zu züchten.

Das bedeutet:

- gewünschte Eigenschaften definieren,
- Eigenschaften getrennt entwickeln,
- unerwünschte Kopplungen beobachten,
- Kombinationen systematisch prüfen,
- riskante Kombinationen vermeiden,
- Kombinationen nur unter bestimmten Bedingungen zulassen.

Der entscheidende Gedanke lautet:

Die Kombination selbst ist eine Sicherheitsentscheidung.

14. Der vorgelagerte Sicherheitsgedanke

Alignment setzt typischerweise bei einem bereits vorhandenen Fähigkeitsraum an.

Das System kann bestimmte Dinge, und anschließend versucht man sicherzustellen, dass es diese Fähigkeiten im gewünschten Sinn verwendet.

Die Zuchtperspektive stellt eine Frage davor:

Warum besitzt das System genau diese Kombination von Fähigkeiten?

Wenn eine riskante Eigenschaftskombination erst einmal vorhanden ist, muss man sie anschließend kontrollieren.

Vielleicht ist es in manchen Fällen sinnvoller, diese Kombination gar nicht erst entstehen zu lassen.

Das entspricht ziemlich genau der Logik der Zucht.

Nicht jede Eigenschaft, die man erzeugen kann, muss mit jeder anderen kombiniert werden.

15. Erste Hypothesen

Aus der Übertragung ergeben sich für mich derzeit folgende Hypothesen.

H1

KI-Risiko hängt nicht nur vom Niveau einzelner Fähigkeiten ab, sondern wesentlich von bestimmten Capability-Kombinationen.

H2

Zwei oder mehrere einzeln nützliche Eigenschaften können gemeinsam eine neue Risikoklasse erzeugen.

H3

Bestimmte Fähigkeiten sollten deshalb nicht zwangsläufig im selben autonomen System vorhanden sein.

H4

Ein Gesamtsystem aus spezialisierten Komponenten kann sehr leistungsfähig sein, ohne dass jede Komponente die gesamte Wirkungskette beherrscht.

H5

Die Rekombination getrennter Fähigkeiten ist selbst ein sicherheitsrelevanter Vorgang.

H6

Capability-Trennung kann schnelle parallele Entwicklung ermöglichen und damit einen Teil des Zielkonflikts zwischen Sicherheit und Entwicklungsgeschwindigkeit entschärfen.

H7

Ein Teil heutiger KI-Sicherheitsprobleme könnte daraus entstehen, dass zu viele Fähigkeiten zuerst zusammengeführt und erst anschließend kontrolliert werden.

16. Der eigentliche Claim

Der stärkste Claim scheint mir derzeit nicht zu sein:

„KI wird gezüchtet statt gebaut.“

Das ist der Einstieg in das Verständnis.

Der eigentliche Claim ist:

Bestimmte KI-Eigenschaften sollten bewusst in getrennten Linien entwickelt und nicht wahllos miteinander kombiniert werden.

Daraus folgt:

Capability-Risiko muss als Problem der Kombination und nicht nur als Problem einzelner Fähigkeiten betrachtet werden.

Und daraus wiederum:

KI-Sicherheit kann teilweise bereits durch die Architektur der Capability-Kombination erzeugt werden.

Und als Rahmen:

Die Organisationsbionik hilft beim Verständnis der Auswirkungen mancher Capability-Kombination.

17. Explizite Claims dieser Primärfassung

1. KI-Entwicklung lässt sich in relevanten Teilen sinnvoll als Selektions- bzw. Zuchtprozess betrachten.
2. Das Bild des „Bauens“ von KI-Systemen kann sicherheitstechnisch relevante Fragen verdecken.
3. KI-Risiko kann aus der Kombination mehrerer Fähigkeiten entstehen.
4. Eine nützliche Fähigkeit plus eine weitere nützliche Fähigkeit muss nicht automatisch eine nützliche oder sichere Kombination ergeben.
5. Bestimmte Capabilities sollten daher in getrennten Entwicklungs- oder Einsatzlinien gehalten werden.
6. Nicht jedes leistungsfähige KI-System muss ein Universalagent sein.
7. Ein Gesamtsystem kann leistungsfähig sein, obwohl einzelne Komponenten bewusst nicht alle Fähigkeiten besitzen.
8. Die Kombination getrennter Fähigkeiten ist selbst sicherheitsrelevant.
9. Die Wirkungskette Wahrnehmung → Zielwahl → Planung → Ausführung → Wirkung muss nicht vollständig in einem einzelnen autonomen System liegen.
10. Capability-Trennung kann möglicherweise dazu beitragen, Geschwindigkeit und Sicherheit besser miteinander zu vereinbaren.
11. Die Gebrauchstierzucht ist ein geeignetes Analogsystem, um diese Frage zu untersuchen.
12. Der Weg von diesem Analogsystem zur KI-Sicherheitsarchitektur ist ein Fall klassischer Analog- bzw. Organisationsbionik.

18. Was noch offen ist

Noch ungeklärt ist insbesondere:

- Welche Capability-Kombinationen tatsächlich besonders riskant sind.
- Welche Fähigkeiten sinnvoll getrennt werden können.
- Welche Trennungen technisch realistisch sind.
- Wann ein Multi-Agent-System dieselben Risiken erzeugt wie ein Universalagent.
- Wie eine formale Prüfung von Capability-Kombinationen aussehen müsste.
- Welche Teile dieser Gedanken bereits in AI Safety, Modular AI, Multi-Agent-Systemen oder Evolutionary Computation vorhanden sind.
- Welche Teile tatsächlich neu aus der organisationsbionischen Übertragung entstehen.

19. Verdichtung

Das Problem lautet:

KI entwickelt sich unter Wettbewerbsdruck immer schneller und bündelt immer mehr Fähigkeiten in denselben Systemen.

Die Analogie lautet:

Bei der Zucht leistungsfähiger Gebrauchstiere werden Eigenschaften nicht beliebig kombiniert. Bestimmte Eigenschaften werden bewusst in getrennten Linien und Rollen gehalten.

Die abstrahierte Erkenntnis lautet:

Nicht nur Eigenschaften selbst, sondern ihre Kombinationen können sicherheitsrelevant sein.

Die Übertragung lautet:

KI-Fähigkeiten getrennt entwickeln, Kombinationen prüfen und nur gezielt rekombinieren.

Die mögliche Lösung lautet:

Nicht langsamer züchten, sondern kontrollierter kreuzen.

Lizenz:

Dieser Text steht unter Creative Commons Attribution 4.0 International (CC BY 4.0). Weitergabe und Bearbeitung, auch kommerziell, sind unter Namensnennung gestattet.

Empfohlene Zitierung:

Martin Hlustik, „KI züchten statt bauen? Ein organisationsbionischer Ansatz zur KI-Sicherheit“, organisationsbionik.de, Version 0.2, 19.09.2026.