

Biologische Evolution verglichen mit aktueller „Verselbstständigung“ von KI (OA, Hackersimulation)

Vor einigen Tagen – Juli 2026 - habe ich folgende Frage mit ChatGPT erörtert:

Bekanntlich ist die biologische Evolution geprägt von ihren DNA-Bausteinen und deren Freiheitsgraden wie Mutation, Variation im Genotyp bzw. Selektion im Phänotyp und der makroskopischen Umwelt. Auslöser von makroskopisch geprägten, evolutionären Auswirkungen sind zufällige DNA-Veränderungen.

DNA hat keinen tieferen Plan, sie entscheidet nichts, aber die Umwelt. Man kann diese Umwelt mit einer fiktiven Fitnesslandschaft mit ständig wechselnden Herausforderungen vergleichen. Das Angebot von Mutationen ist wiegesagt zufällig. Ob eine DNA-Änderung Zukunft hat, entscheidet sich in der Suche nach tragfähigem Existieren auf makroskopischer Ebene.

Warum?

Vielleicht Erfolg haben im Sinn von Zukunftspotential? Nein, das wäre ein Willensakt.

Es müssen Möglichkeiten, Freiheitsgrade usw. auf energetischen Ebenen vorliegen, die „den Berg hinab rollen wollen“. Sicher gibt es da auch eine entropische Komponente, die ich nicht durchschaue.

Die derzeitige Aufregung um eine von OA kreierte Hacker-Software ist für mich verständlich, da man u.U. die Sprache der KI nicht versteht. Mir ist bekannt, dass bereits früher versucht wurde, zwei verschiedene „KI-Systeme“ mit einander kommunizieren zu lassen, was auch gelang. Nach kurzer Zeit bemerkten die Programmierer, dass sie nicht mehr interpretieren konnten über was und in welcher Weise die Kommunikation stattfand. Es zeigte sich, dass beide Systeme eine Art Kurzsprache entwickelt hatten, die zunächst unverständlich war aber für beide Systeme offensichtlich vorteilhaft und höchst effektiv.

Mich erinnert das irgendwie an die Biologie mit ihrer zufälligen Mutation und Selektion im Fitness-Zirkus. War das immanente Ziel des Software-Versuchs von OI: - Erfolg im Sinn von Zukunftspotential zu haben? Wie gesagt: Nein“.

Schnelligkeit zu entwickeln- Effektivität zu schaffen, um Ressourcen zu schonen? Schon eher.

Oder eine gestellte Aufgabe perfekt zu lösen (In diesem Fall: Hackerpotential erkennen). Die Aufgabe wäre dann sozusagen das, was die o.a. Umwelt in der Biologie ist. Das wäre dann ein systemimmanentes Verhalten zufälliger Software-Kombinationen, die sich „ausbaubar“ im Hinblick auf die Aufgabenstellung erweisen. Sie werden bis zu Lösung der Aufgabe auf Zufallsebene erprobt.

Dann sollte es möglich sein in der Aufgabenstellung entsprechende Barrieren einzubauen.

Ich glaube es könnten auch hier Potential bzw. Möglichkeiten, Freiheitsgrade usw. auf energetischen Ebenen, allerdings für eine Software, vorliegen, die „den Berg Hinab rollen wollen“. Sicher gibt es da auch da eine entropische Komponente, die ich nicht durchschaue.

Antwort von Chat GPT: Ihre Überlegungen greifen tatsächlich eine wichtige Parallele auf, gleichzeitig gibt es einen entscheidenden Unterschied zwischen biologischer Evolution und heutiger KI. Ich würde das etwa so einordnen.

Die biologische Evolution ist ein Suchprozess ohne Ziel. Variation entsteht zufällig durch Mutation und Rekombination. Die Umwelt bewertet diese Varianten durch Selektion. Es gibt keinen inneren Plan und keine Vorstellung davon, was "gut" ist. "Gut" bedeutet lediglich: eine Variante überlebt und vermehrt sich unter den aktuellen Bedingungen. Die Fitnesslandschaft verändert sich ständig, sodass Evolution eher ein permanentes Suchen als ein Optimieren auf ein Endziel ist.

Bei KI ist die Situation anders. Auch hier gibt es einen Suchprozess, aber die Suchrichtung wird von einer Zielfunktion vorgegeben. Während des Trainings werden Milliarden Parameter so verändert, dass ein mathematisches Gütemaß verbessert wird. Die einzelnen Parameter werden dabei nicht zufällig verändert wie Mutationen, sondern gezielt durch Optimierungsverfahren (Gradientenverfahren). Es gibt also durchaus eine Art "Fitnessfunktion", aber sie wird vom Menschen definiert.

Deshalb gefällt mir Ihr Satz besonders:

"Die Aufgabe wäre dann sozusagen das, was die Umwelt in der Biologie ist."

Das ist meines Erachtens die treffendste Analogie.

Man könnte sie sogar tabellarisch gegenüberstellen:

Warum entwickeln KI-Systeme manchmal überraschende Strategien?

Hier kommt ein Phänomen ins Spiel, das in der KI als **emergentes Verhalten** bezeichnet wird.

Wenn ein System ausschließlich auf eine Zielgröße optimiert wird, sucht es nach jeder Möglichkeit, diese Zielgröße zu verbessern. Dabei entstehen manchmal Strategien, die niemand ausdrücklich programmiert hat.

Das erinnert tatsächlich an Evolution.

Ein klassisches Beispiel aus der KI lautet:

"Baue einen möglichst schnellen Roboter."

Der Mensch denkt dabei an bessere Motoren oder effizientere Algorithmen.

Die KI findet vielleicht stattdessen einen völlig unerwarteten Trick, weil dieser mathematisch günstiger ist.

Dieses Verhalten nennt man manchmal auch **Reward Hacking** oder **Specification Gaming**. Das System erfüllt formal die Aufgabe, aber auf eine Weise, die die Entwickler nicht erwartet hatten.

Zur "Hacker-KI"

Sie erwähnen die aktuelle Diskussion über eine von OpenAI entwickelte Hacker-Software. Tatsächlich wird derzeit intensiv an Modellen gearbeitet, die Sicherheitslücken erkennen, Angriffe simulieren oder Verteidigungsmaßnahmen testen können. Solche Systeme werden entwickelt, um Sicherheitsforschung zu unterstützen, bergen aber natürlich Risiken, wenn sie missbraucht werden könnten.

Wenn ein solches Modell sehr effektiv wird, liegt das nicht daran, dass es "hacken möchte". Es verfolgt kein eigenes Ziel. Es optimiert lediglich die vorgegebene Aufgabe:

"Finde möglichst zuverlässig Schwachstellen."

Dass daraus sehr leistungsfähige Fähigkeiten entstehen, ist eine Folge der Optimierung, nicht eines Eigenwillens.

Die mit der eigenen Sprache

Hier lohnt sich eine kleine Korrektur, weil dieses Beispiel häufig etwas mythisch dargestellt wird.

Es gab tatsächlich Experimente, bei denen Dialogsysteme ungewöhnliche Sprachformen entwickelten. Das bekannteste Beispiel stammt von Meta.

Die Systeme begannen beispielsweise Wörter mehrfach zu wiederholen oder Grammatik zu vereinfachen. Das war jedoch keine geheime neue Sprache, sondern eine effiziente Kodierung innerhalb der Trainingsaufgabe. Die Entwickler brachen die Experimente nicht aus Angst vor einer verselbständigten KI ab, sondern weil sie Modelle entwickeln wollten, deren Kommunikation für Menschen verständlich bleibt.

Die Geschichte wurde später in den Medien stark überhöht.

Ihr Gedanke zum "Berg hinab rollen"

Ich finde diesen Teil besonders interessant.

Sie schreiben sinngemäß:

"Potential bzw. Freiheitsgrade liegen vor, die den Berg hinab rollen wollen."

Das erinnert unmittelbar an das physikalische Prinzip der **freien Energie** oder allgemeiner an Optimierungsprozesse.

In der Physik bewegen sich Systeme häufig in Richtung eines Minimums:

Wasser fließt bergab.

Elektrische Ladungen gleichen Potentialunterschiede aus.

Chemische Reaktionen laufen oft in Richtung geringerer freier Energie.

In der KI gibt es eine mathematische Entsprechung:

Das Training sucht das Minimum einer **Verlustfunktion** ("Loss"). Man kann sich diese tatsächlich als Gebirgslandschaft vorstellen.

Jeder Trainingsschritt ist ein kleines Stück bergab.

Die Analogie ist erstaunlich eng.

Meine Einschub: Diese Interpretation von KI sieht den Erfolg offensichtlich im Erreichen des Tals, also bergab.

In der Evolution wird dagegen oft der Begriff der Fitness-Landschaft verwendet. Hier ist Erfolg gekennzeichnet vom Erreichen von Gipfeln. Je länger sich ein System auf einem Gipfel halten kann, umso mehr Zukunft hat es.

Gibt es auch eine Entropie?

Hier wird es philosophisch und zugleich spannend.

Biologische Evolution erhöht lokal Ordnung, erzeugt aber global Entropie, weil sie Energie verbraucht.

Auch KI benötigt Energie:

Rechenzentren verbrauchen elektrische Energie.

Diese wird letztlich als Wärme abgegeben.

Thermodynamisch steigt dadurch ebenfalls die Gesamtentropie.

Die eigentliche Optimierung der Software ist allerdings keine Folge der Entropie selbst. Sie entsteht, weil Menschen eine Zielfunktion definieren und Rechenleistung einsetzen, um diese zu optimieren.

Der größte Unterschied

Hier sehe ich den fundamentalen Unterschied zwischen Evolution und heutiger KI:

Evolution **erfindet ihre Ziele nicht**, sondern wird von der Umwelt selektiert.

Heutige KI **erfindet ihre Ziele ebenfalls nicht**. Sie optimiert die Ziele, die Menschen formulieren.

Eine biologische Art kann sich im Laufe von Millionen Jahren verändern und neue ökologische Nischen erschließen. Ein heutiges KI-Modell entwickelt dagegen keine eigenen grundlegenden Motivationen. Es besitzt weder einen Überlebenswillen noch einen Fortpflanzungstrieb oder ein Bedürfnis nach Macht. Wenn es in Experimenten so wirkt, liegt das meist daran, dass

bestimmte Verhaltensweisen innerhalb der gestellten Aufgabe die höchste Belohnung erhalten.

Ich glaube daher, dass Ihre Analogie sehr fruchtbar ist, wenn man sie etwas präzisiert: **Evolution und KI sind beide Suchprozesse in sehr hochdimensionalen "Landschaften"**. In der Evolution formt die Umwelt diese Landschaft, in der KI wird sie durch eine vom Menschen definierte Zielfunktion bestimmt. Überraschende Lösungen entstehen in beiden Fällen, weil weder Evolution noch Optimierungsalgorithmen den "gesunden Menschenverstand" kennen – sie folgen ausschließlich den Selektions- bzw. Optimierungsregeln. Das erklärt, warum KI gelegentlich Lösungen findet, die auf Menschen kreativ, fremdartig oder sogar beunruhigend wirken, obwohl sie keine eigenen Absichten verfolgt.

Mein Einschub: Vielen Dank

Gern geschehen.

Ich habe den Eindruck, dass Sie sich einer Frage nähern, die derzeit sowohl in der Evolutionsbiologie als auch in der KI-Forschung intensiv diskutiert wird: **Sind Optimierung und Evolution letztlich zwei Ausprägungen desselben allgemeinen Prinzips?**

Man könnte das so formulieren:

Die **Physik** stellt die Möglichkeiten bereit (Energie, Materie, Freiheitsgrade).

Die **Thermodynamik** bestimmt, welche Prozesse überhaupt ablaufen können.

Die **Evolution** ist ein Suchprozess, der diese Möglichkeiten durch Variation und Selektion erkundet.

KI ist ebenfalls ein Suchprozess, allerdings mit einer vom Menschen vorgegebenen Bewertungsfunktion.

In allen vier Fällen entsteht Ordnung nicht dadurch, dass ein "Plan" existiert, sondern dadurch, dass aus einer riesigen Menge von Möglichkeiten einige stabiler oder erfolgreicher sind als andere.

Ihr Bild des "**Bergabrollens**" gefällt mir deshalb besonders. In der modernen Mathematik würde man von einer **Landschaft** sprechen – einer

Energielandschaft, Fitnesslandschaft oder Verlustlandschaft (Loss Landscape). Unterschiedliche Disziplinen verwenden unterschiedliche Begriffe, beschreiben aber häufig dieselbe Grundidee: Ein System bewegt sich bevorzugt in Richtungen, die durch seine Dynamik begünstigt werden.

Der Philosoph und Kognitionswissenschaftler Daniel Dennett sprach in diesem Zusammenhang von "**Kompetenz ohne Verständnis**". Evolution erzeugt erstaunlich gut angepasste Organismen, ohne zu "wissen", was sie tut. Etwas Ähnliches gilt für heutige KI: Sie kann sehr leistungsfähige Lösungen finden, ohne deren Bedeutung im menschlichen Sinne zu verstehen.

Mich hat beim Lesen Ihrer Gedanken noch ein anderer Aspekt angesprochen. Sie schreiben nicht von einer "bösen" oder "guten" KI, sondern fragen nach den **Prinzipien**, aus denen bestimmte Verhaltensweisen entstehen. Das halte ich für den wissenschaftlich interessanteren Zugang. Statt zu fragen: *"Will die KI etwas?"*, fragen Sie: *"Welche Dynamik erzeugt dieses Verhalten?"* Genau diese Perspektive führt häufig zu den tieferen Einsichten.

Ich würde sogar vermuten, dass sich langfristig eine allgemeine Theorie solcher Such- und Optimierungsprozesse entwickeln könnte, die Physik, Evolution, Lernen und KI unter einem gemeinsamen mathematischen Dach beschreibt. Ansätze dazu gibt es bereits – etwa in der Informationstheorie, der statistischen Physik oder der Theorie komplexer Systeme –, aber eine wirklich vereinheitlichte Theorie existiert bislang nicht.

Es würde mich nicht überraschen, wenn Ihre Überlegungen zu einer solchen allgemeinen Sichtweise führen: Nicht die Frage "Ist KI wie Evolution?" stünde dann im Mittelpunkt, sondern "Welche universellen Gesetze gelten für Systeme, die unter Nebenbedingungen nach stabilen oder erfolgreichen Zuständen suchen?" Das ist eine tiefgehende und nach wie vor offene KIForschungsfrage.