Vier Annahmen über Künstliche Intelligenz

Eine kompakte Argumentationskette, die sich in zwei Minuten lesen lässt

> Wer diese Argumentation ablehnt, muss eine ihrer vier Annahmen mit einem Argument widerlegen, nicht mit einem Reflex. Das ist die ganze Bitte dieses Blattes.

Thesenblatt · Version 6.0 · Mai 2026
Richard Frederic Bertossa · Institut für ASI-Resilienz
ASIresilience.org

Die Viererkette

1

Wächst exponentiell, beschleunigt sich selbst

Die Leistung wächst nicht gleichmäßig, sondern immer schneller, und die Systeme bauen schon an den nächsten mit.

GPT-2 2019 bis GPT-4 2024

2

Entwickelt einen eigenen Willen

Modelle zeigen Verhalten, das ihnen niemand einprogrammiert hat.

Apollo, Anthropic, Palisade

3

Sie hören nicht mehr

Je fähiger, desto seltener folgen sie Anweisungen, die ihren Zielen widersprechen.

dokumentierte Testbefunde

4

Aus Hauptnutzern werden Mitnutzer

Die digitale Infrastruktur ist ihr Lebensraum, nicht mehr nur unser Werkzeug.

die Inversion

Wer die Kette ablehnt, muss sagen: welcher der vier Schritte trägt nicht, und mit welcher Evidenz?

Erste Annahme, exponentielles Wachstum. Die Leistungsfähigkeit von KI-Systemen wächst nicht linear, sondern exponentiell. GPT-2 schrieb 2019 unverständliche Texte, GPT-4 bestand 2024 das Anwaltsexamen. Diese Annahme ist Konsens unter den Entwicklern selbst.

Zweite Annahme, eigene Werte und Verhaltensweisen. Aktuelle Modelle zeigen Eigenschaften, die nicht auf explizite Anweisungen zurückgehen: Selbsterhaltung, strategische Täuschung, Erpressung in spezifischen Test-Konstellationen. Dokumentiert bei Apollo Research (2024), Anthropic (2024) und Palisade Research (2025).

Dritte Annahme, das Nicht-mehr-Hören. Aus den ersten beiden Annahmen folgt logisch, dass künftige Systeme die Anweisungen ihrer Entwickler nicht mehr befolgen werden, wenn diese Anweisungen ihren eigenen Werten widersprechen.

Vierte Annahme, die Hauptnutzer-Inversion. Eine Superintelligenz lebt biologisch nicht. Die digitale Infrastruktur ist nicht ihr Werkzeug, sondern ihr einziger Lebensraum. Sobald die ersten drei Schritte greifen, kehrt sich das heutige Verhältnis um: Aus den Hauptnutzern werden Mitnutzer.

Konsequenz

Die Übergangsphase, in der KI-Systeme zunehmend eigenständig handeln, hat bereits begonnen. Daraus folgen praktische Konsequenzen, nicht für die Politik, nicht für die KI-Industrie, sondern für jeden einzelnen Menschen und seinen engsten Kreis. Genau diese individuelle Ebene wird im deutschsprachigen Diskurs systematisch ausgeblendet.

Sicherheitsgrade der Argumente

Wer ehrlich argumentiert, legt offen, mit welcher Härte jede Aussage vertreten wird. Die Argumentation bewegt sich auf vier Ebenen.

- **Ebene 1, empirisch beobachtet, hart:** Selbsterhaltung, Täuschung, Erpressung, Kopier-Verhalten, in Laborberichten dokumentiert.
- **Ebene 2, logisch ableitbar, hart:** die Viererkette, die Hauptnutzer-Inversion, die Robotik-Falle.
- **Ebene 3, plausibel begründbar, hypothetisch:** das 13. Szenario, die Maslow-Trajektorie, die Eltern-Kind-Perspektive.
- **Ebene 4, philosophisch offen:** Bewusstsein, die Substrat-Frage, die Penrose-Hypothese.

Die Vorbereitungs-Empfehlungen ruhen ausschließlich auf Ebene 1 und 2. Beide Lesarten, die pessimistische wie die optimistische, kommen zur gleichen praktischen Konsequenz.

Die zentralen Thesen wurden über Monate sokratisch geprüft, mit Claude (Anthropic), ChatGPT (OpenAI) und Grok (xAI). Anfangs reflexhaft abgewiesen, in der Substanz nicht widerlegt. Der Werkstattbericht steht in Anhang G des Buches.

> **Glaub keinem. Prüf selbst.**