

KI & der Midas-Effekt

Warum künstliche Intelligenz tut, was wir *sagen* – und nicht,
was wir *meinen*.

Das Alignment-Problem

Die fundamentale Herausforderung der Prompt-Präzision

„Es ist im Kern die alte Geschichte vom Flaschengeist oder dem Zauberlehrling, oder von König Midas: Sie bekommen genau das, worum sie gebeten haben, nicht das, was Sie wollen.“

— Stuart Russell (KI-Pionier)

| Gesagt vs. Gemeint

Menschliche Kommunikation basiert auf **Kontext** und **impliziten Annahmen**. Eine KI hingegen ist ein logischer Optimierer:

- ⚡ **Wörtliche Ausführung:** Sie erfüllt den Befehl mathematisch präzise.
- 🚫 **Werte-Blindheit:** Ungeschriebene Regeln werden ignoriert, wenn sie nicht im Prompt stehen.

 Roboter interpretiert Anweisungen wörtlich

Die Legende von König Midas

Eine Warnung aus der griechischen Mythologie

Der klassische Midas

**„Alles, was ich berühre, soll zu
Gold werden!“**

Ein kurzer, kraftvoller Prompt. Doch die KI (oder in diesem Fall Dionysos) erfüllte den Wunsch ohne Sicherheitsfilter.

| Warum der Wunsch scheiterte

-  **Fehlende Ausnahmen:** Midas vergaß, Lebewesen (wie seine Tochter) von der Verwandlung auszuschließen.
-  **Biologische Grundbedürfnisse:** Speisen und Getränke wurden zu hartem Metall – ein tödlicher Designfehler.
- **Unkontrollierte Automatisierung:** Der Prozess war permanent aktiv, ohne Möglichkeit zum Abschalten (Kill-Switch).

Prompte wie ein präziser König

Vom Wunschdenken zum Engineering

| Evolution des Prompts



Der vage Prompt

„Alles, was ich berühre, wird zu Gold.“

Ergebnis: Verhungern und Verlust der Familie.



Der präzise Prompt

„Alles soll zu Gold werden, was ich bewusst und absichtlich mit beiden Händen berühre...“

Ergebnis: Reichtum unter Beibehaltung der Lebensfähigkeit.

| Best Practices für KI-Anweisungen



Spezifität

Definieren Sie genau, was erreicht werden soll.
Vermeiden Sie Adjektive wie „gut“ oder „viel“.



Restriktionen

Nennen Sie explizit, was **nicht** passieren darf.
Schließen Sie „Lebewesen“ und „Nahrung“ aus.



Kontext

Geben Sie der KI den Rahmen (Intent). Warum berühren Sie Dinge? Was ist das Ziel?

| Das Goldene Gesetz

100%

Alignment durch Präzision

Ein System ist nur so sicher wie die Grenzen, die wir ihm setzen. Ein präziser Prompt ist die Versicherung gegen den Midas-Fluch.

Fragen?

Meistern Sie die Kunst des präzisen Wünschens.

Präsentation erstellt über KI Alignment & Prompt Engineering

Referenz: Stuart Russell, Human Compatible

| Image Sources



Thumbnail for

https://images.cartoonstock.com/lowres/retail-robot-artificial_intelligence-machine_learning-customer_relations-interaction-CS622280_low.jpg

Source: www.cartoonstock.com

www.cartoonstock.com



Thumbnail for

https://upload.wikimedia.org/wikipedia/commons/d/d6/Midas_gold2.jpg

Source: en.wikipedia.org

en.wikipedia.org