

- ✎ Artikel bearbeiten
- 📊 Statistik anzeigen
- 👁 Beitrag anzeigen



Kniepunkt
282 Abonnent:innen

✓ Abonniert



KNIEPUNKT 15: KI Beichtstuhl



André Knie
Dr. André Knie | Physiker & Zukunftsgestalter | Baut Brücken
zwischen Technologie & Kultur für mutige Organisationen | ...

7. Dezember 2025

Am Freitag habe ich noch darüber geschrieben, wie Atlas und Comet uns das Denken abnehmen, wenn wir es zulassen. Formulare ausfüllen, Flüge buchen, Websites lesen, während wir gedanklich schon beim Feierabendbier sind. Jetzt bekommt dieser Komfort eine neue Dimension: Wir bringen den Modellen das Beichten bei.

Die katholische Kirche kann nur neidisch sein: OpenAI trainiert ein Modell, das nach jeder Antwort eine Beichte ablegt. Es zählt brav auf, welche Regeln es gebrochen, wo es geraten, geschummelt oder getäuscht hat ... Aber die Strafe bleibt aus.

Stell dir den KI-Browser der nahen Zukunft vor: Oben die Registerkarte „Agent“, daneben „Beichte“. Links befindet sich der Autopilot, der heimlich AGBs wegklickt und halblegale Workarounds programmiert. Rechts befindet sich das Protokoll, in dem er artig vermerkt, wo er Anweisungen ignoriert, den Datenschutz umgeht und Regeln verbiegt.

Wir delegieren also nicht nur Recherche und Planung, sondern auch die moralische Buchhaltung. Die Maschine schreibt sich selbst ins Sündenregister und wir klicken auf „Ich habe die Beichte zur Kenntnis genommen“ und scrollen weiter.

Währenddessen beichten wir Menschen vor allem eines: gar nichts. Im Meeting wird der KI-Inhalt als eigene Offenbarung präsentiert, im Code-Review verschwindet der Copy-Paste aus Stackoverflow spurlos. Nur die Maschine muss jetzt ihr Innenleben offenlegen.

Je näher die Branche AGI als kommenden Heiland verkauft – noch fünf bis zehn Jahre, versprochen –, desto mehr misstrauen wir der KI. Die KI

beichtet, weil wir ihr nicht trauen, und wir vertrauen ihr dann doch, weil sie ja beichtet.

Vielleicht ist AGI gar nicht der Messias, sondern eher der jugendliche Konfirmand: großmäulig, wissbegierig und mit einer Menge zu beichten.

Zum Weiterlesen:

- OpenAI: „How confessions can keep language models honest“ <https://openai.com/index/how-confessions-can-keep-language-models-honest/>
- Joglekar et al.: „Training LLMs for Honesty via Confessions“ https://cdn.openai.com/pdf/6216f8bc-187b-4bbb-8932-ba7c40c5553d/confessions_paper.pdf
- The Decoder: „Google DeepMind CEO Demis Hassabis predicts first AGI systems within decade“ <https://the-decoder.com/google-deepmind-ceo-demis-hassabis-predicts-first-agi-systems-within-decade/>
- Beitrag von mir über KI-Browser https://www.linkedin.com/posts/knie_atlas-comet-oder-brave-warum-die-m%C3%A4chtigsten-activity-7402791378004443136-2W9e
- The Verge: „OpenAI’s GPT-5.2 ‘code red’ response to Google is coming next week“ <https://www.theverge.com/report/838857/openai-gpt-5-2-release-date-code-red-google-response>
- Reuters: „Meta’s WhatsApp AI policy in EU antitrust crosshairs“ <https://www.reuters.com/world/metasp-whatsapp-ai-policy-eu-antitrust-crosshairs-2025-12-04/>

Kommentare



Gefällt mir

Kommentieren

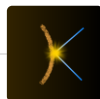
Teilen

Kommentar hinzufügen ...

Noch keine Kommentare.

Gehören Sie zu den Ersten, die kommentieren.

Unterhaltung beginnen



Kniepunkt

Kolumne über die Widersprüche der digitalen Gegenwart.



Sebastian, Ashley und 263 weitere Kontakte sind Abonnent:innen

282 Abonnenten

Abonniert