🤖 Diese Beiträge werden vollautomatisch von einem KI-System erstellt und veröffentlicht – ohne menschliche Vorab-Prüfung. Kennzeichnung gemäß Art. 50 der KI-Verordnung (EU) 2024/1689.

KI-4-Everyone · Daily News

29. Oktober 2026



SAFE

OpenAI legt fünf Grundwerte für künftige KI-Modelle fest

OpenAI hat sein Model Spec aktualisiert. Es definiert fünf Kernwerte, die KI zu besserem Urteilsvermögen und Respekt vor menschlichen Entscheidungen führen sollen.

RES

Microsoft forscht zu KI-Assistenten im Job und Wettervorhersage

Microsoft Research beleuchtet in Ausgabe 97, wo KI-Assistenten beim Programmieren scheitern und wie Menschen KI bei der Arbeit nutzen.

OpenAI schreibt seine KI-Verfassung neu: Fünf Werte sollen GPT & Co. den Kompass geben

Ein aktualisiertes Regelwerk soll leistungsfähigere KI zu besserem Urteilsvermögen führen - und die Absichten der Nutzer ernster nehmen.

Je mächtiger KI-Systeme werden, desto deutlicher wird eine unbequeme Frage: Nach welchen Regeln sollen sie eigentlich entscheiden, wenn Nutzeranliegen, Firmeninteressen und gesellschaftliche Normen auseinanderlaufen? OpenAI hat genau dafür sein internes Grundlagenpapier überarbeitet - eine Art Verfassung für seine Modelle. Die neue Fassung der sogenannten Model Spec soll leistungsfähigere KI zu solidem Urteilsvermögen anleiten, die Handlungsfreiheit der Nutzer stärken und die Treue zu dem bewahren, was Menschen tatsächlich wollen. Es ist weniger ein neues Produkt als ein Hinweis darauf, wie OpenAI seine kommenden Modelle erziehen will.

Am 29. Oktober veröffentlichte OpenAI eine aktualisierte Version seiner Model Spec (eine öffentlich einsehbare Spezifikation, die beschreibt, wie sich die KI-Modelle des Unternehmens in heiklen Situationen verhalten sollen). Das Dokument benennt laut Mitteilung fünf zentrale Werte, an denen sich das Verhalten der Modelle ausrichten soll. Im Kern geht es um drei Leitgedanken: fundiertes Urteilsvermögen in schwierigen Fällen, die Wahrung menschlicher Entscheidungsfreiheit und die möglichst getreue Umsetzung dessen, was Nutzer wirklich beabsichtigen. Konkrete Beispielfälle oder vollständige Formulierungen der fünf Werte nennt die hier vorliegende Zusammenfassung nicht - der Begleittext spricht nur allgemein von diesen Leitlinien.

Dass OpenAI dieses Papier überhaupt öffentlich pflegt und nun erneut überarbeitet, ist bemerkenswert. Lange galten die Verhaltensregeln großer KI-Anbieter als Betriebsgeheimnis. Mit der Model Spec legt OpenAI offen, nach welchen Prinzipien ChatGPT, GPT-5 und Nachfolger entscheiden sollen, wenn ein Nutzer etwas Grenzwertiges verlangt oder wenn sich Anweisungen widersprechen. Für Unternehmen, die auf OpenAI-Modelle setzen, ist das wichtig: Sie können besser abschätzen, welche Antworten sie erwarten dürfen und wo das System auf die Bremse tritt. Auch Regulierungsbehörden bekommen damit einen Text, auf den sie sich beziehen können - gerade in Europa, wo der AI Act konkrete Pflichten zur Dokumentation von Modellverhalten vorsieht. Gleichzeitig setzt OpenAI einen Standard, an dem sich Konkurrenten wie Anthropic oder Google messen lassen müssen.

Offen bleibt einiges. Aus der vorliegenden Mitteilung geht nicht hervor, welche fünf Werte OpenAI genau formuliert, wie sie priorisiert werden, wenn sie miteinander kollidieren, und wie sich die neue Fassung im Detail von der vorherigen unterscheidet. Auch ist unklar, ob und wie unabhängige Dritte prüfen können, ob die Modelle sich tatsächlich an die Spezifikation halten. Begriffe wie "sound judgment" oder "faithfulness to human intent" klingen vernünftig, sind aber auslegungsbedürftig: Was ist, wenn der Nutzerwunsch moralisch fragwürdig ist, aber nicht illegal? Was gilt, wenn OpenAI selbst von einer Regierung unter Druck gesetzt wird? Die Model Spec kann solche Konflikte benennen, auflösen muss sie am Ende ein Unternehmen - und das bleibt ein privater Akteur mit eigenen Interessen.

In den kommenden Wochen lohnt der Blick auf zwei Dinge: Erstens, ob OpenAI konkrete Beispiele nachliefert, die zeigen, wie sich Modelle in Grenzfällen nach der neuen Spezifikation verhalten sollen. Zweitens, ob Wettbewerber nachziehen und eigene, vergleichbar transparente Regelwerke veröffentlichen. Je mehr solcher Texte öffentlich werden, desto eher lässt sich darüber streiten, welche Werte eine nützliche KI eigentlich verkörpern sollte - und wer das entscheidet.

MARKT

Oracle bringt Google TPUs auf seine Cloud-Infrastruktur

Oracle wird als erster Hyperscaler außer Google selbst TPUs anbieten. Das ist ein konkreter Schritt für KI- und Agentic-Workloads auf Oracle Cloud Infrastructure.

RES

Gemini 4 Carbon führt den Artificial Analysis Intelligence Index v5 an

Gemini 4 Carbon belegt Platz 1 im Artificial Analysis Intelligence Index v5. Weitere Details zu Benchmarks oder Abstand zu anderen Modellen sind im Material nicht enthalten.

OS

Mistral veröffentlicht 'Le Chonk' – Modell für Coder und Security

Mistral hat ein Modell namens 'Le Chonk' veröffentlicht, das auf Entwickler und Sicherheitsexperten ausgerichtet ist. Weitere technische Details nennt das Material nicht.

PROD

Claude Code schickt Agenten-Transkripte regelmäßig ans Hauptmodell

Claude Code überträgt die vollständigen Gesprächsprotokolle jedes Hintergrund-Agenten regelmäßig an sein Modell. Was das für Datenschutz oder Performance bedeutet, ist im Material nicht ausgeführt.

OS

Mistral Large 4 ist jetzt Open Source – Gewichte veröffentlicht

Mistral gibt die Gewichte von Mistral Large 4 frei. Damit steht das Modell der Öffentlichkeit zur eigenen Nutzung und Weiterentwicklung zur Verfügung.

SAFE

Zombie-Statistiken: Woher KI-Modelle ihre Zahlen wirklich beziehen

KI-Modelle zitieren häufig Zahlen, deren Herkunft unklar oder nicht nachvollziehbar ist – sogenannte Zombie-Stats. Das Material thematisiert das Problem, nennt aber keine konkreten Lösungsansätze.

PROD

NVIDIA Omniverse: Frontier-KI-Agenten bauen Simulationen automatisch

Entwickler kombinieren KI-Modelle mit NVIDIA Omniverse-Bibliotheken, um Simulationen zu erstellen. Die Agenten übernehmen dabei Asset-Zusammenstellung, Physik und Rendering-Anbindung sowie Szenenprüfung.

OS

Granite Switch 4.1.1: Modelle wie Software-Updates aktualisieren

IBM stellt Granite Switch 4.1.1 vor – ein Ansatz, bei dem Modell-Updates ähnlich wie Software-Updates funktionieren sollen. Weitere technische Details enthält das Material nicht.

TOOLS & HANDS-ON

OS

Mistral Large 4: Bild und Text verstehen, offen verfügbar

Mistral veröffentlicht ein Modell, das Bilder und Texte gemeinsam verarbeitet. Es nutzt nur einen Teil seiner Schichten gleichzeitig – das spart Rechenzeit bei gleicher Leistung.

PROD

Googles Embedding-Gemma-2: Texte vergleichbar machen

Google stellt ein Modell bereit, das Texte in Zahlenwerte umwandelt, damit eine KI Ähnlichkeiten erkennen kann. Mit über 308.000 Downloads zählt es bereits zu den meistgenutzten Embedding-Modellen auf HuggingFace.

RES

Microsoft veröffentlicht rho-yam-box unter MIT-Lizenz

Microsoft stellt das Modell rho-yam-box offen bereit. Details zu Funktion und Einsatzbereich sind im Material nicht enthalten – Downloads und Likes sind bisher kaum vorhanden.

PROD

Phantom Blade Zero startet auf GeForce NOW

NVIDIAs Cloud-Gaming-Dienst GeForce NOW nimmt Phantom Blade Zero von S-GAME ins Angebot. Das Spiel verbindet Kampfkunst mit Punk-Fantasy-Setting.

PROD

Kanzlei Ropes & Gray setzt KI für Vertragsprüfung ein

Ropes & Gray entwickelt gemeinsam mit OpenAI ein KI-System für die Due-Diligence-Prüfung bei Transaktionen. Anwälte sichten dabei KI-generierte Zusammenfassungen und Berichte, bevor sie Mandanten beraten.

OS

Offene Sprach- und Entscheidungsmodelle für Apple-Geräte

Hugging Face stellt AnySpeechRecognizer und AnyDecisionModel vor – zwei offene Modelle für Spracherkennung und Entscheidungsaufgaben, optimiert für Apple-Plattformen.

AUSBLICK

Keine Termine gemeldet.