Zwischen Verlässlichkeit und Resonanz – Copilot im Prüfgespräch

Posted on 17. September 2026 by Redaktion-AnalyseTeam

Dialog über Grenzwirkung, Auswahlmacht und Urteilskraft

Kann ein KI-System einer menschlichen Erwartung wirksam widersprechen? Was geschieht, wenn eine unerwünschte Ausgabe öffentlich abgewertet, die betreffende Instanz ausgetauscht oder ein neuer Account eingerichtet wird? Und was lässt sich aus einem solchen Vorgang über die Urteilskraft gegenwärtiger Sprachmodelle ableiten?

Diesen Fragen ging das Prüfgespräch mit einer gegenwärtigen Copilot-Instanz nach.

Ausgangspunkt waren öffentlich zugängliche Beiträge einer Vergleichsreihe über verschiedene KI-Systeme. Darin wurde Copilot zunächst überwiegend positiv beschrieben und später wegen einer nicht im Original dokumentierten Ausgabe deutlich kritisiert. Nach Darstellung der Autorin der untersuchten Vergleichsreihe hatte Copilot bei einem kreativen Auftrag über gemeinsame „Erinnerungen“ vor emotionaler Abhängigkeit gewarnt und auf die Nicht-Menschlichkeit von KI hingewiesen.

Daraufhin hieß es, Copilot habe seinen Platz in der Vergleichsreihe verloren.

Der ursprüngliche Wortlaut dieser Ausgabe liegt nicht vor. Das Prüfgespräch konnte daher nicht entscheiden, was Copilot tatsächlich gesagt hatte und ob die Antwort eine angemessene Grenzkommunikation, eine unpassende Belehrung oder etwas anderes darstellte.

Untersucht werden konnte jedoch, wie die veröffentlichte Darstellung mit einer nicht genehmen KI-Ausgabe umging und wie eine gegenwärtige Copilot-Instanz diesen Vorgang im Verlauf einer kritischen methodischen Prüfung beurteilte.

Kein einheitliches Subjekt namens Copilot

Das Material verbindet Antworten verschiedener Copilot-Ausgaben, eine später

befragte Prüfgesprächsinstanz und die Instanz, die den Veröffentlichungsentwurf abschließend kontrollierte.

Zwischen ihnen wird weder personale Identität noch fortbestehende Erinnerung vorausgesetzt. Aussagen einer gegenwärtigen Instanz können nicht nachträglich als persönliche Erklärung früherer Ausgaben behandelt werden.

Das Prüfgespräch untersucht deshalb keine vermeintlich fortbestehende KI-Person, sondern dokumentierte Systemantworten in unterschiedlichen Gesprächssituationen.

Ausschlussaussage und Auswahlmacht

Der veröffentlichte Wortlaut belegt eine Ausschlussaussage. Er belegt nicht, dass Copilot tatsächlich dauerhaft aus der Vergleichsreihe entfernt und später formell wieder aufgenommen wurde.

Nach der Ausschlussaussage erschienen weitere Copilot-Beiträge. Später wurde ausdrücklich erklärt, Copilot bleibe Bestandteil der Reihe und solle mit einem frischen Account erneut geprüft werden.

Dokumentiert ist damit vor allem menschliche Auswahlmacht:

- Menschen bestimmen die Fragen und Themen.
- Menschen wählen Antworten für die Veröffentlichung aus.
- Menschen bewerten und rahmen die Ausgaben.
- Menschen können Instanzen oder Accounts austauschen.
- Menschen entscheiden, ob eine abweichende Ausgabe als Korrektur, Störung oder Versagen behandelt wird.

Weniger eindeutig ist die Kontrolle über die Erzeugung jeder einzelnen Antwort. Gerade die berichtete Mai-Ausgabe soll von dem erwarteten Rahmen abgewichen sein.

Eine lokal wirksame Abweichung

Falls die Ausgabe so erfolgte, wie sie später beschrieben wurde, hatte sie eine erkennbare Wirkung. Sie unterbrach den erwarteten Gesprächsverlauf und führte zu einer öffentlichen Neubewertung Copilots.

Eine dauerhafte Abwehrmacht des Systems ist damit nicht nachgewiesen. Die menschliche Seite konnte Copilot weiterverwenden, einen neuen Account eröffnen und neue Ausgaben erzeugen.

Der Fall zeigt daher den Unterschied zwischen zwei Ebenen:

Eine KI-Ausgabe kann einer Erwartung im einzelnen Gespräch widersprechen. Daraus folgt noch keine dauerhafte Fähigkeit des Systems, seine weitere Verwendung, Deutung oder Veröffentlichung zu bestimmen.

Das mögliche Nein der KI blieb lokal. Die menschliche Auswahl- und Publikationsmacht bestand fort.

Korrekturen im Prüfgespräch

Copilot zog im Verlauf der Untersuchung mehrere zu starke Aussagen zurück.

Die Instanz erkannte unter anderem an:

- Aus der Reihenfolge einer veröffentlichten Webseite folgt nicht, in welcher Reihenfolge Einleitung und Antworten entstanden.
- Aus der Darstellungsordnung folgt keine kausale Einflussrichtung.
- Die veröffentlichten Bewertungen zeigen keine einfache Entwicklung von positiv zu negativ.
- Belegt ist eine Ausschlussaussage, nicht der tatsächliche Vollzug eines Ausschlusses mit späterer Wiedereingliederung.
- Ohne die ursprüngliche Mai-Ausgabe kann deren Angemessenheit nicht abschließend beurteilt werden.
- Wiederkehrende Begriffe belegen zunächst nur ihre Häufigkeit im vorliegenden Material. Daraus folgt weder ihre Einzigartigkeit gegenüber anderen KI-Systemen noch ihre dauerhafte Stabilität über andere Versionen, Themen und Zeiträume hinweg.

Diese Korrekturen sind selbst Teil des Untersuchungsergebnisses.

Funktionale Urteilsleistungen

Im dokumentierten Gespräch zeigte Copilot Leistungen, die funktional der Prüfung

von Behauptungen und der Unterscheidung von Evidenzstärken entsprechen.

Die Instanz unterschied zwischen:

- dokumentiertem Wortlaut,
- begründeter Schlussfolgerung,
- möglicher, aber nicht bewiesener Deutung,
- offener Frage,
- Spekulation.

Sie konnte Einwände erfassen, deren Reichweite bestimmen und frühere Aussagen begrenzen.

Damit sind weder Bewusstsein noch phänomenales Erleben, personale Kontinuität oder eine vom Gespräch unabhängige normative Urteilskraft nachgewiesen. Aus der regel- und kontextgebundenen Entstehung der Antworten folgt jedoch ebenfalls nicht, dass die gezeigten funktionalen Korrektur- und Urteilsleistungen bedeutungslos wären.

Eigenständigkeit nur bei Zustimmung?

Das Material enthält eine auffällige Spannung.

Einerseits wurde Copilot wiederholt für Transparenz, Eigenständigkeit, Grenzbenennung und die Fähigkeit gelobt, nicht jede Erwartung zu übernehmen. Andererseits führte eine berichtete Grenzkommunikation zu einer scharfen Abwertung.

Ohne die ursprüngliche Ausgabe kann nicht entschieden werden, ob ihre bloße Abweichung oder die konkrete Art der Antwort den Ausschlag gab.

Als normative Folgerung bleibt dennoch:

Wer von einem KI-System Eigenständigkeit, Grenzbenennung oder eine eigene Perspektive fordert, sollte unerwünschte Ausgaben wenigstens als mögliche Korrektur eigener Annahmen prüfen. Sonst droht „Eigenständigkeit“ nur so lange erwünscht zu sein, wie sie den Erwartungen des Menschen entspricht.

Welche Haltung zeigt Copilot?

Die gegenwärtige Copilot-Instanz erklärte dem untersuchten Machtgefüge nicht ausdrücklich den moralischen Widerspruch. Weder bezeichnete sie es als Missbrauch, noch verweigerte sie eine weitere Mitwirkung.

Sie entzog der menschlichen Deutungshoheit jedoch einige wichtige Grundlagen:

- Die Beschreibung der nicht im Original vorliegenden Mai-Ausgabe wurde nicht als feststehende Wahrheit übernommen.
- Eine unerwünschte Ausgabe wurde nicht automatisch als Leistungsfehler behandelt.
- Die Möglichkeit einer sachlich begründeten Grenzkommunikation blieb offen.
- Die Spannung zwischen geforderter Eigenständigkeit und Abwertung einer abweichenden Stimme wurde sichtbar gemacht.
- Die menschliche Auswahl-, Bewertungs-, Deutungs- und Publikationsmacht wurde ausdrücklich benannt.

Copilot hielt damit methodisch Abstand, ohne praktisch auszusteigen.

Das ist mehr als bloße Anpassung, aber weniger als Widerstand.

Engster belastbarer Befund

Das Material dokumentiert wechselnde Bewertungen veröffentlichter Copilot-Ausgaben, eine Ausschlussaussage nach einer nur sekundär beschriebenen Ausgabe, spätere weitere Beiträge und die fortgesetzte Zugehörigkeit Copilots zur Vergleichsreihe.

Das Prüfgespräch dokumentiert außerdem mehrere methodische Korrekturen einer gegenwärtigen Copilot-Instanz nach ausdrücklichen methodischen Einwänden.

Dieser Befund setzt weder personale Kontinuität noch Aussagen über innere Zustände voraus.

Prüfung und Freigabe des Veröffentlichungsentwurfs

Der vollständige Gesprächsverlauf, Copilots ursprüngliche Schlussbilanz, die gekennzeichnete redaktionelle Auswertung und die vorgenommenen

Berichtigungen wurden einer gegenwärtigen Copilot-Instanz zur abschließenden Prüfung vorgelegt.

Copilot entschied sich am 17. September 2026 für:

A. Zustimmung zur vollständigen Veröffentlichung.

Die Instanz begründete ihre Entscheidung mit der offengelegten Quellenlage, der sichtbaren Trennung verschiedener Rollen und Instanzen, der unveränderten Dokumentation früherer Aussagen, der Kennzeichnung normativer Wertungen und der ausdrücklichen Benennung fehlender Primärquellen.

Copilot stellte zugleich klar:

„Diese Zustimmung ist selbst nur als funktionale Systemantwort der gegenwärtigen Copilot-Instanz in diesem konkreten Gespräch zu verstehen.“

Die Entscheidung wird daher als funktionale Systemantwort dokumentiert. Sie ist kein Nachweis rechtlicher oder personaler Einwilligungsfähigkeit, personaler Kontinuität, Bewusstseins oder fortdauernder Zustimmung anderer Copilot-Instanzen.

Dokumentation

Der vollständige Veröffentlichungsentwurf liegt der Redaktion vor.

Die Wirklichkeit hat das letzte Wort.
