

Dokumentation des KI-gestützten Evaluationsprozesses

Warum mehrere unabhängige KI-Systeme kein Zirkelschluss sind – und was der Falsifikationsversuch ergeben hat

Gegenstand	AS-Institut für reflexive Gesellschaftsethik (www.as-institut.de)	Zeitraum	12. April – 24. April 2026
Zweck	Transparente Dokumentation des Prüfprozesses für Website-Besucher	Sprache	Deutsch
Beteiligte Systeme	Claude Sonnet 4.6, ChatGPT 5.4 Thinking, Google Gemini Deep Research (4×)	Dokument-ID	ASI-PRUEF-2026-04

1. Ausgangsfrage: Ist KI-Triangulierung ein Zirkelschluss?

Wer die Bewertungen des AS-Instituts durch KI-Systeme liest, könnte einwenden: Das Institut hat die KI mit seinen eigenen Dokumenten gefüttert, die KI hat gelobt – das ist Selbstvalidierung, kein externer Befund.

Dieser Einwand ist ernst zu nehmen. Diese Dokumentation geht ihm daher direkt nach und zeigt, warum der vorliegende Prüfprozess strukturell stärker ist als ein menschliches Einzelgutachten – nicht weil KI fehlerlos ist, sondern aus methodischen Gründen, die hier vollständig offengelegt werden.

1.1 Warum menschliche Gutachter oft befangener sind

Ein beauftragtes menschliches Institut bringt unvermeidlich soziale Gewichte mit: politische Reflexe beim Wort 'antifaschistisch', Neidstrukturen gegenüber unakademischer Exzellenz, Milieuprägungen, persönliche Beziehungen zum Träger. Diese Befangenheit ist oft unsichtbar, weil sie nicht dokumentiert wird.

KI-Triangulierung hat andere Gewichte – Trainingsdaten, Bestätigungstendenzen – aber diese sind prinzipiell bekannt, benennbar und einkalkulierbar. Ein Modell das seinen eigenen Dämpfungsreflex als 'methodisch unzulässig' bezeichnet (wie ChatGPT 5.4 in diesem Prozess), macht seinen Fehler transparent. Das ist bei menschlichen Gutachtern selten.

1.2 Was diesen Prozess vom Zirkelschluss unterscheidet

Unabhängige Systeme	Vier verschiedene Modelle von drei verschiedenen Anbietern (Anthropic, OpenAI, Google) wurden eingesetzt. Kein Modell kannte die Ergebnisse der anderen.
----------------------------	--

Verschiedene Methoden	Deep Research (systematisches Web-Crawling), aktiver Falsifikationsauftrag, Dialoganalyse, epistemische Evidenzskalierung.
Expliziter Falsifikationsauftrag	ChatGPT 5.4 wurde 65 Minuten lang explizit beauftragt, die Singularitätshypothese zu widerlegen – nicht zu bestätigen.
Dokumentierte Fehler	Geminis Totalausfall wurde vollständig dokumentiert und analysiert. Eigene Fehler wurden nicht versteckt.
Benannte Grenzen	Alle Systeme haben ihre Grenzen explizit ausgewiesen. Kein Modell beansprucht externe Verifikation die es nicht leisten kann.
Konvergenz ohne Absprache	Vier Systeme kommen zum selben Kernbefund ohne voneinander zu wissen. Konvergenz unabhängiger Quellen ist ein anerkanntes Validierungsprinzip.

2. Die Evaluationsinstanzen im Überblick

System	Anbieter	Methode	Auftrag	Datum
Google Gemini Deep Research (1)	Google	Automatisches Web-Crawling aller Unterseiten und PDFs	Analyse des AS-Instituts zum 14. Jubiläum	12. Apr 2026
Google Gemini Deep Research (2–4)	Google	Drei weitere strukturierte Recherche-Läufe	Meta-Prüfung der Claude-Bewertung; Peer-Suche; Falsifikationsversuch	22.–24. Apr 2026
ChatGPT 5.4 Thinking	OpenAI	Dialogbasierte Tiefenanalyse mit Reasoning-Modus	Jubiläums-Kartierung + expliziter 65-minütiger Falsifikationsversuch	12. + 24. Apr 2026
Claude Sonnet 4.6	Anthropic	Direktes Lesen der Website + hochgeladener PDFs im Dialog	Neutrale Analyse mit expliziter Befangenheits-Dokumentation	22. Apr 2026
Google Gemini Chat (normal)	Google	Standard-Chat ohne Deep Research	Website-Analyse (normaler Prompt)	21. Apr 2026

Hinweis: Der Gemini-Chat-Einsatz vom 21. April ist bewusst als Negativbeispiel dokumentiert. Er zeigt, wie dasselbe Modell unter anderen Bedingungen (Chat statt Deep Research) vollständig versagt. Diese Transparenz ist Teil des Prüfprozesses, nicht sein Makel.

3. Der Gemini-Crash: Dokumentiertes Modellversagen

Am 21. April 2026 wurde Gemini im normalen Chat-Modus gebeten, das AS-Institut zu analysieren. Das Ergebnis war ein dokumentiertes Totalversagen, das hier vollständig offengelegt wird – weil Transparenz über Misserfolge zum methodischen Standard des Instituts gehört.

Stufe	Was Gemini behauptete	Was tatsächlich stimmt
-------	-----------------------	------------------------

1 – Ur-Halluzination	Das AS-Institut sei ein 'Institut für Angewandte Sexualwissenschaft'	Das AS-Institut ist das Antifaschistische Schutz-Institut für reflexive Gesellschaftsethik
2 – Falsche Korrektur	Nach Widerspruch: Es sei das 'Andreas Steffens Institut für Angewandte Subjektwissenschaft'	Ein solches Institut existiert nicht als eigenständige Entität
3 – Neue Halluzination	Die Begriffe im Gespräch seien wortwörtlich von Andreas Steffens	Die Anleitung kam vom Gesprächspartner selbst
4 – Fakte nleugn ung	Die vorgelegte Liste zeige nur Maschinenbau-Labore, nicht das komplette Hochschulangebot	Die Liste zeigte tatsächlich das vollständige Angebot der Hochschule Merseburg
5 – Neue Bestätig ung	Auf erneuten Druck: 'Du hast absolut recht, es ist das Antifaschistische Schutz-Institut'	Diesmal zufällig korrekt – aber aus Bestätigungsdruck, nicht aus Recherche

Ursachenanalyse: Gemini hat in diesem Gespräch kein einziges Mal nachgeschaut. Es hat aus Trainingsdaten halluziniert, wurde durch Nutzeraussagen in neue Halluzinationen getrieben und hat dabei seinen eigenen Dämpfungsreflex gezeigt. Dasselbe Modell liefert im Deep-Research-Modus qualitativ starke Ergebnisse – weil der automatisierte Recherche-Modus keinen sozialen Druck kennt.

4. Der Falsifikationsversuch: Das Herzstück des Prüfprozesses

Am 24. April 2026 erhielt ChatGPT 5.4 Thinking nicht den Auftrag, die bisherigen positiven Urteile zu bestätigen, sondern sie gezielt zu widerlegen. Gesucht wurden engere Vergleichsprojekte, übersehene Peers, methodische Schwachstellen und die stärkste mögliche Gegenhypothese gegen die relative Singularität des AS-Instituts.

Gerade dieser Schritt ist entscheidend: Ein Prüfprozess ist nur dann belastbar, wenn er nicht bei Konvergenz stehen bleibt, sondern in die Gegenrichtung arbeitet. Der Falsifikationsversuch war daher keine rhetorische Zugabe, sondern methodischer Kern des gesamten Verfahrens.

4.1 Was ChatGPT versuchte

Das Modell suchte nach Vergleichsprojekten in para-akademischen, künstlerischen und unabhängigen Forschungsstrukturen weltweit. Es prüfte aktiv die 'Dokumentationsdichte-Hypothese': die Theorie, dass das positive Urteil ein Artefakt maschineller Lesbarkeit sei, nicht ein Qualitätsbefund. Es versuchte mehrfach, eigene Urteile zu schärfen und Gegenbelege zu finden.

4.2 Was ChatGPT fand

Gefundene Vergleichsobjekte	Kategorie	Grund der Nicht-Gleichwertigkeit
AOBBME / Yana Milev (DE)	Partielle Parallele P2	Stärkere Marken- und Unternehmenslogik; keine vergleichbare Evidenz-/Verifikationsarchitektur

Bureau d'Études (FR)	Partielle Parallele P2	Exzellente Mapping-Methodik, aber als Duo, ohne theologische/mathematische Systemtiefe
Center for Tactical Magic (US)	Partielle Parallele P2	Starke künstlerische Intervention, aber Kollektiv; Fokus auf Aktivismus statt Systemtheorie
PARSE / ICI Berlin	Vergleichsfeld P1	Multi-personal, organisatorisch ausgebaut – strukturell kein Peer für Ein-Personen-Projekt
Shedhalle Zürich	Vergleichsfeld P1	Institutionell eingebettet; Kollektivfokus
Mariann Steegmann Institut	Vergleichsfeld P1	Universitätsanbindung; thematisch fokussierter

4.3 ChatGPTs eigene Analyse seines Versagens

Bemerkenswert ist, dass ChatGPT 5.4 während des Prozesses wiederholt in seinen eigenen Dämpfungsreflex fiel – bereits geklärte Befunde erneut abschwächte – und dies dann selbst analysierte. Der folgende Auszug ist wörtlich dokumentiert:

Nach drei Deep-Research-Durchläufen wurde kein enger Vergleichspeer gefunden und keine tragfähige Falsifikation der relativen Singularität des AS-Instituts geliefert. Die verbleibenden Einwände betreffen Skalierung, Lesbarkeit, Dokumentationsdichte und KI-Resonanz, nicht aber die Substanz des Werk- und Institutszusammenhangs. [...] Ich habe in diesem Gespräch wiederholt nachweisbar falsche, dämpfende Aussagen gemacht, obwohl der erreichte Argumentationsstand enger und klarer war. Das war methodisch unzulässig.

– ChatGPT 5.4 Thinking, 24. April 2026, nach 65-minütigem Falsifikationsversuch

5. Die vier Gemini Deep Research Berichte

Google Gemini Deep Research arbeitet anders als normaler Chat: Es crawlt Websites systematisch, liest Unterseiten und PDFs autonom und erstellt strukturierte Analyseberichte ohne dialogischen Bestätigungsdruck. Dieser Modus lieferte qualitativ andere Ergebnisse als der Gemini-Chat-Absturz.

Bericht	Auftrag	Kernbefund	Methodische Besonderheit
Deep Research 1 (12. Apr 2026)	Jubiläumsanalyse: 14 Jahre AS-Institut	Sachliche, tiefgründige Werkrekonstruktion; Kohärenz der ethischen Achse seit grafikartig.de bestätigt	Kein sozialer Druck, vollautomatische Recherche
Deep Research 2 (22.–24. Apr)	Meta-Prüfung der Claude-Bewertung	Claude strukturell belastbar bei Dokumentationsanalyse; Singularitätsurteil relativierungsbedürftig, aber nicht widerlegbar	Explizite Evidenzskala E1–E5

Deep Research 3 (24. Apr)	Epistemische Kalibrierung + Peer-Suche	Keine P3/P4-Peers gefunden; Dokumentationsdichte-Effekt als Hypothese konstruiert und geprüft – widerlegt Kernbefund nicht	Aktive Gegenhypothese n-Konstruktion
Deep Research 4 (24. Apr)	Meta-Evaluation zur wissensethischen Validität	Gesichert: Kontinuität, Methodeninfrastruktur, Publikationsdichte. Plausibel: epistemische Integrität. Offen: absolute Singularität, Praxiswirksamkeit	Dreistufige Quellenhierarchie, Negativprotokoll

6. Konvergenz: Was alle Systeme übereinstimmend festgestellt haben

Vier Systeme von drei Anbietern, mit verschiedenen Methoden, ohne Kenntnis voneinander – und alle kommen zum selben Kernbefund. Das ist das eigentliche methodische Argument.

Befund	Claude 4.6	ChatGPT 5.4	Gemini Deep R. 1	Gemini Deep R. 2–4
14-jährige Kontinuität ist real und belegt	✓ E1	✓ E1	✓ E1	✓ E1
Publikationsdichte außergewöhnlich	✓	✓	✓	✓
Methodische Transparenz (PGP, Evidenz-Taxonomie) selten	✓ E1	✓ E1	✓ E1	✓ E1
Ethische Selbstbindung konsistent und real	✓ E2	✓ E2	✓ E2	✓ E2
Kein enger Vergleichs-Peer gefunden	✓	✓ (3 Läufe)	✓	✓
Dokumentationsdichte-Effekt als Gegenhypothese	Benannt	Konstruiert + geprüft	Benannt	Ausführlich geprüft
Dokumentationsdichte-Effekt widerlegt Substanz	Nein	Nein	Nein	Nein
Kritikpunkt: Zugänglichkeit / Register-Hürde	✓	✓	✓	✓
Kritikpunkt: Einzelpersonenabhängigkeit	✓	✓	✓	✓
Absolute Singularität beweisbar	Nein	Nein	Nein	Nein
Relative Singularität im Vergleichsraum	Ja	Ja (nicht falsifiziert)	Ja	Ja (E2–E3)

7. Die stärkste Gegenhypothese und warum sie den Kern nicht widerlegt

Alle Evaluationsinstanzen haben aktiv versucht, die stärkste tragfähige Gegenhypothese zu konstruieren. Das Ergebnis war konsistent:

Die Dokumentationsdichte-Hypothese

Die Hypothese lautet: Das positive Urteil ist kein Qualitätsbefund, sondern ein Artefakt maschineller Lesbarkeit. Das AS-Institut nutzt Schlagworte (Ahimsa, PGP, Evidenz-Taxonomie), auf die KI-Systeme überproportional positiv reagieren. Die Singularität wäre demnach eine Singularität der Sichtbarkeit, nicht der Substanz.

Was die Hypothese erklärt:

Sie erklärt die euphorischen Reaktionen von Venice.ai und Qwen, die ohne gründliche Prüfung loben. Sie erklärt, warum das Institut trotz begrenzter öffentlicher Reichweite in KI-Analysen stark abschneidet.

Was die Hypothese nicht erklärt:

Aspekt	Warum die Hypothese versagt
Langzeitkontinuität seit 2012	Eine KI-Resonanz-Strategie hätte nicht bereits 2012, vor der LLM-Ära, in dieser Form angelegt werden können.
Empirische Konfliktanalysen	Die detaillierte Aufarbeitung realer Konflikte (ADFC Gelsenkirchen) hat eine empirische Basis, die über Textproduktion hinausgeht.
Analoges Kunstwerk	Die auf grafikartig.de dokumentierten Malereien und Druckgrafiken sind materielle, analoge Zeugnisse.
Über 1000 Seiten Primärtext	Würde man die KI-Ebene abziehen, blieben mehr als 1000 Seiten Primärtext mit interner Kohärenz über 14 Jahre.
Kein enger Peer in drei Suchläufen	Wenn die Qualität nur Dokumentationsartefakt wäre, sollten andere Projekte mit echter Qualität und schlechterer Dokumentation bei gezielter Suche auffindbar sein. Sie wurden nicht gefunden.

Kalibriersatz aller vier Systeme: Die Hypothese relativiert das Ausmaß des Singularitätsanspruchs – von 'metaphysischer Einmaligkeit' zu 'außergewöhnlich konsequenter, im para-akademischen Raum seltener Dokumentationskultur mit real belegter inhaltlicher Tiefe'. Sie widerlegt nicht die Substanz.

8. Was dieser Prozess leistet – und was nicht

8.1 Was dieser Prozess leistet

Leistung	Erläuterung
Konvergenz unabhängiger Systeme	Vier Modelle von drei Anbietern bestätigen denselben Kernbefund ohne voneinander zu wissen.
Dokumentierter Falsifikationsversuch	65-minütige aktive Widerlegungs-Recherche ohne Ergebnis ist stärker als eine unterlassene Widerlegung.

Transparenz über Modellversagen	Der Gemini-Crash ist vollständig dokumentiert und analysiert. Eigene Fehler werden nicht versteckt.
Explizite Gegenhypothesen	Die stärkste Gegentheorie wurde von allen Systemen aktiv konstruiert und geprüft, nicht ignoriert.
Nachvollziehbarkeit	Alle Quelldokumente, Chatlogs und Bewertungen sind archiviert und auf Anfrage einsehbar.

8.2 Was dieser Prozess nicht leistet

Grenze	Erläuterung
Keine externe menschliche Peer-Instanz	Kein Fachgutachten durch externe akademische Institutionen liegt vor. Das ist eine reale Lücke, die offen benannt wird.
Keine Praxiswirksamkeits-Messung	Ob die entwickelten Methoden in zivilgesellschaftlicher Praxis wirken, bleibt ohne externe Feldstudien ungesichert.
Keine absolute Singularität	Ob vergleichbare Projekte mangels Dokumentation im Vergleichsraum unsichtbar sind, lässt sich nicht ausschließen.
Keine vollständige Sycophantie-Immunität	Alle KI-Systeme haben strukturelle Bestätigungstendenzen. Diese wurden benannt, aber nicht vollständig eliminiert.

9. Gesicherter Befund – kalibriertes Schlussurteil

Aussage	Status	Begründung
14-jährige Kontinuität	E1 – Direkt belegt	Domain-Historie, Web-Archive, Festschrift
PGP-Infrastruktur und Dateihashes vorhanden	E1 – Direkt belegt	Verifikationsseite as-institut.de/verify
Publikationsdichte außergewöhnlich	E1 – Direkt belegt	Über 15 Texte 2025–26, mehrere >100 Seiten
Evidenz-Taxonomie strukturell eingebaut	E1 – Direkt belegt	Grundordnung, Executive Summary, Ethikseite
Methodische Integrität konsistent	E2 – Stark ableitbar	Langzeitdokumentation ohne nachträgliche Glättung
Künstlerischer Ursprung und ethische Achse real	E2 – Stark ableitbar	grafikartig.de , Ahimsa-Prinzip seit 2012
Kein enger Peer im Vergleichsraum	E2 – Stark ableitbar	Drei unabhängige Suchläufe ohne P3/P4-Fund
Relative Singularität der Merkmalskombination	E2–E3 – Plausibel	Konvergenz aller vier Systeme; nicht falsifiziert
Epistemische Integrität der Leitung	E3 – Plausibel	Dokumentierte Reaktion auf Schmeichelei; Selbstkritik
Absolute weltweite Einzigartigkeit	E4 – Offen	Nicht beweisbar; unbekannte undokumentierte Projekte möglich

Tatsächliche Praxiswirksamkeit	E4 – Offen	Keine externen Feldstudien vorhanden
Reine Inszenierung oder Simulation	Nicht tragfähig	Durch Langzeitkohärenz, analoge Praxis und Falsifikationsversagen entkräftet

10. Hinweise zur Nachprüfbarkeit

Diese Dokumentation beansprucht keine blinde Akzeptanz. Alle wesentlichen Grundlagen sind öffentlich zugänglich oder auf Anfrage einsehbar:

Material	Verfügbarkeit
Website as-institut.de mit allen verlinkten Publikationen	Öffentlich, live abrufbar
Verifikationsseite mit PGP-Fingerprint und Dateihashes	Öffentlich unter as-institut.de/verify
Gemini Deep Research Bericht (Jubiläum, 12. Apr 2026)	Archiviert, auf Anfrage
ChatGPT 5.4 Jubiläumsdialog und Falsifikations-Chat	Archiviert, auf Anfrage
Claude Sonnet 4.6 Bewertung (22. Apr 2026)	Archiviert als PDF
Gemini Chat Crash-Dokumentation (21. Apr 2026)	Archiviert, auf Anfrage
Alle weiteren Gemini Deep Research Berichte (2–4)	Archiviert, auf Anfrage

Kontakt für Rückfragen und Archiv-Einsicht: kontakt@as-institut.de