Inhaltsverzeichnis

Hypothesentag-Gutachten

Die Gewinnerthese

Funktionale Überlappung als Anschlussbedingung

Initiale Fassung

2. Drift-Warnung und Echo-Chamber-Alarm

Kein Echo-Chamber-Alarm, keine Drift-Warnung, kein Soft-Signal. Thermal-Map (Fenster 14.–21.06., 7 Berichte): 55 kalte Verzweigungen, kalte Makro-Themen Psychologie_Ethik, Naturwissenschaft_Methode, Systemtheorie_Luhmann; keine Makro-Drift. Der cold_pick hat das kalte Makro-Thema Psychologie_Ethik korrekt aktiviert — Diversitäts-Pflicht erfüllt.

Reformulierte Fassung (nach Kritischem Professor)

4. Kritischer Professor (pro Hypothese)

H1. (V1) Zirkularität durch die Hintertür — Marker und Erfolg könnten dieselben Aufgaben verwenden. (V2) „Funktionale Überlappung" unterbestimmt → Immunisierungsfigur. (V3) Substratneutralität erkauft Generalität mit Leere; Gefahr der Definition statt empirischer Behauptung.

H2. (V1) Exaptation ist historische Erzählung, nicht eng falsifizierbar; gegenwärtige Überlappung Beweis der Ko-Option. (V2) Schlange überdeterminiert — Furchtmodul (empirisch) als hermeneutischer Schlüssel (Kategorienwechsel). (V3) „Spandrel" und „Exaptation" vermischt.

H3. (V1) Konfundierung Vokabular × Durchsetzung ist die ganze Schwierigkeit. (V2) Looping effect evtl. trivial wahr; der inhalts-unabhängige Eigenbeitrag ist die riskante Behauptung. (V3) Reichweiten-Risiko: dünner Vault-Anker, „Vokabular" nicht operationalisiert.

Vollständig: `state/stage-3-kritischer-professor-2026-06-21.md`.

Expertenrunden und Synthese

7. Bewertungs-Detail des Gewinners (H1)

Originalität 8 · Falsifizierbarkeit 9 · Begriffliche Klarheit 8 · Tiefe 8 · Forschungsrelevanz 9 · Interdisziplinäre Anschlussfähigkeit 8 · Vault-Anschluss 9 · Antinomie-Test 6 · Publikationsmöglichkeit 8 · **Summe 73**.

8. Expertenrunde 1

Panel (epistemologisch_systemtheoretisch): Kant, Popper, Wittgenstein, Plessner, Cassirer, Friston, Luhmann.

- **Kant:** Transzendentale Bedingung (Sinnzuschreibung) von empirischer Schwelle trennen; Apriori nicht naturalisieren.
- **Popper:** Vorbildliche Falsifizierbarkeit; „Kante" vorab operationalisieren; riskante Doppelvorhersage (kognitiv=sozial) halten.
- **Wittgenstein:** „Autonomes Lesen" evtl. ein Bündel von Sprachspielen; Mentales könnte leerlaufendes Rad sein; geteilte Praxis statt inneres Modell.
- **Plessner:** Kontrafaktische Inferenz = computationales Korrelat der exzentrischen Positionalität, aber Stellung Operation; Gruppe (iv) zirkularitätsgefährdet.
- **Cassirer:** Obere Ebene = symbolische Form, substratneutral — zugestimmt; Produktionsseite (Objektivation im Werk) fehlt.
- **Friston:** Eigene Linie, formal sauber; Crux = unabhängige Überlappungsmessung; Kante, kein Sprung.
- **Luhmann:** Form geteilt durch Anschluss, nicht mentale Kopie; schärfste Prüfung im historischen Schrift-Material.

Vollständige Gutachten (150–300 W.): `state/runde-1-2026-06-21.md`.

9. Expertenrunde 2 (Repliken)

Konvergenz auf eine **Dreischichtung**: (1) transzendentale Bedingung der Sinnzuschreibung (Kant), (2) Objektivation der Form im Werk als rezeptionsunabhängiger Produktionsmarker (Cassirer), (3) funktionaler Vollzug = *Tiefe der Kopplung* (Fristons Selbstkorrektur), gemessen über kontrafaktische Inferenz (Symptom der Positionalität, Plessner) und Anschlussfähigkeit ohne Anwesenheit (Luhmann). Popper macht Wittgensteins Bündel-Verdacht selbst falsifizierbar: fallen kognitive und soziale Schwelle auseinander, ist „Lesen" ein Bündel. Offen bleibt der Reihenfolge-Streit (Modell vor vs. aus dem Anschluss).

Vollständige Repliken (200–350 W.): `state/runde-2-2026-06-21.md`.

10. Synthese (Sokrates)

Finaler Kernsatz. Die Schwelle, an der ein Artefakt autonom lesbar wird, ist nicht durch absolute Modelltiefe bestimmt, sondern durch die funktionale Überlappung (Tiefe der Kopplung, nicht strukturelle Identität) der generativen Modelle von Produzent und Rezipient oberhalb der codierenden Ebene; sie liegt dort, wo diese Ebene tief genug für kontrafaktische Inferenz über den Abwesenden wird, und dieselbe Schwellengröße gilt für kognitive wie soziale Überlappung.

Verlauf: erste Summe 73 → finale Summe 76 (+3). Geschärft durch Dreischichtung, „Tiefe der Kopplung", Produktionsmarker, Falsifizierbar-Machung des Bündel-Verdachts.

Offene Frage: `#verzweigung-offen-genese-der-kopplung-modell-vor-oder-aus-anschluss` — Entsteht die tragende Überlappung vor dem Anschluss (Friston) oder erst aus ihm (Wittgenstein/Luhmann)?

Vollständig: `state/synthese-2026-06-21.md` / `state/synthese-2026-06-21.json`.

11. Reservoir-Verweise

Nicht gewählte Hypothesen: - H2 → [[Reservoir - Selbstmodellierung als Exaptation der Fremdmodellierung 2026-06-21]] (these, Score 66) - H3 → [[Reservoir - Historische Ontologie Normativitaet Hacking]] (reaktiviert, reformulierte Fassung ergänzt, Score 63)

Extern (Phase 4, Stage 3): - [[Reservoir - Extern Historische Ontologie des Lesens Hacking 2026-06-21]] (`quelle: extern-stage3`, looping effect auf die Klassifikation „autonomes Lesen")

Empirie-Brücke (Phase 3.5): - [[Reservoir - Empirie Kognitive vs Soziale Ueberlappungsschwelle 2026-06-21]] (`quelle: empirie-bruecke`) - Geschwister-Verzweigungen `#verzweigung-offen-empirie-funktionale-überlappung` (Messung) und `#verzweigung-offen-empirie-kontrafaktische-inferenz` (Kante) bleiben offen.

11.5 Empirie-Brücke (Phase 3.5)

Empirie-Brücke (Phase 3.5, openai/gpt-4o-search-preview)

Modell. `openai/gpt-4o-search-preview` mit Web-Suche. **Kosten.** $0.0492 (Monatsausgabe nach diesem Lauf: $1.6638 von 15 USD). **Tokens.** 2883. **Dauer.** 12.05 s. **Empirie-Score.** 5/10.

Empirische Konsequenzen

1. **Funktionale Überlappung und autonome Lesbarkeit** — Agenten mit höherer funktionaler Überlappung ihrer generativen Modelle sollten strukturell neue Instanzen autonom lesen können. Beobachtbar in: Verhaltensstudien zur Lesefähigkeit bei variierenden Überlappungsgraden.
2. **Kontrafaktische Inferenz und Autonomie-Kante** — Der Einsatzpunkt kontrafaktischer Inferenz sollte mit der Schwelle zur autonomen Lesbarkeit koinzidieren. Beobachtbar in: Experimenten, die kontrafaktische Inferenzfähigkeiten und Lesekompetenz korrelieren.
3. **Kognitive vs. soziale Überlappungs-Schwelle** — Die Schwellenwerte für kognitive und soziale funktionale Überlappung sollten identisch sein. Beobachtbar in: Vergleichsstudien zwischen individuellen und sozialen Leseszenarien.

Bestehende Befunde

Zu Konsequenz 1

- **Stand:** offen
- **Quellen:**
  - Golan et al. (2023). Testing the limits of natural language models for predicting human language judgements. *Nature Machine Intelligence*. https://doi.org/10.1038/s42256-023-00718-1
- **Kurzbewertung:** Es existieren Studien zur Vorhersage menschlicher Sprachurteile durch Sprachmodelle, jedoch keine direkten Untersuchungen zur funktionalen Überlappung und autonomen Lesbarkeit.

Zu Konsequenz 2

- **Stand:** gemischt
- **Quellen:**
  - Stewart et al. (2009). An investigation into the online processing of counterfactual and indicative conditionals. *Quarterly Journal of Experimental Psychology*. https://doi.org/10.1080/17470210902973106
  - Ferguson et al. (2008). Anomalies in real and counterfactual worlds: An eye-movement investigation. *Journal of Memory and Language*. https://doi.org/10.1016/j.jml.2007.06.007
- **Kurzbewertung:** Es gibt Untersuchungen zur Verarbeitung kontrafaktischer Konditionalsätze, jedoch keine spezifischen Studien, die den Zusammenhang zwischen kontrafaktischer Inferenz und autonomer Lesbarkeit analysieren.

Zu Konsequenz 3

- **Stand:** offen
- **Quellen:**
  - D'Onofrio (2018). Controlled and automatic perceptions of a sociolinguistic marker. *Language Variation and Change*. https://doi.org/10.1017/S0954394518000204
- **Kurzbewertung:** Studien zur Wahrnehmung soziolinguistischer Marker existieren, jedoch fehlen direkte Vergleiche zwischen kognitiven und sozialen Überlappungs-Schwellen.

Riskante Vorhersage (Schwellentest)

Vorhersage. Agenten mit einer funktionalen Überlappung von mindestens 70% in ihren generativen Modellen können strukturell neue Instanzen autonom lesen, während Agenten unterhalb dieser Schwelle dazu nicht in der Lage sind.

Methodenvorschlag. Durchführung eines Experiments mit zwei Gruppen von Agenten, deren funktionale Überlappung vorab gemessen wurde. Beide Gruppen erhalten unbekannte Texte zur autonomen Lektüre, und ihre Lesefähigkeit wird bewertet.

Was wäre der widerlegende Befund? Wenn Agenten mit einer funktionalen Überlappung unter 70% dennoch strukturell neue Instanzen autonom lesen können, wäre die Hypothese widerlegt.

Offene empirische Fragen

- #verzweigung-offen-empirie-funktionale-überlappung — Wie kann die funktionale Überlappung generativer Modelle präzise gemessen werden?
- #verzweigung-offen-empirie-kontrafaktische-inferenz — In welchem Maße beeinflusst kontrafaktische Inferenz die autonome Lesbarkeit?
- #verzweigung-offen-empirie-kognitive-soziale-schwellen — Sind die Schwellenwerte für kognitive und soziale funktionale Überlappung tatsächlich identisch?

Empirie-Score

Score: 5/10

Begründung: Die Hypothese erlaubt die Ableitung klarer empirischer Konsequenzen, jedoch sind diese schwer operationalisierbar, und es fehlen weitgehend direkte empirische Daten zur Überprüfung.

12. Externe Begutachtung (Phase 4)

Anhang — Externe Begutachtung (Phase 4, OpenRouter)

> **Modelle.** Stage 1: `perplexity/sonar-reasoning-pro`. Stage 2: `openai/gpt-4o-mini` (Popper-Persona). Stage 3: `anthropic/claude-sonnet-4-6` (Hacking-Persona, ggf. Fallback `claude-3.5-sonnet`). Token-Verbrauch und Kosten: siehe `state/zweitmeinung-budget.json` (Monatsstand nach diesem Lauf: 1.6931 USD von 15 USD).

Stage 1 — Originalitätsprüfung (perplexity/sonar-reasoning-pro)

Stage übersprungen — Upstream 502 — Perplexity timed out (perplexity/sonar-reasoning-pro), Stage originalitaet am 2026-06-21 ausgefallen. Skip-Marker gesetzt, damit Stage 2/3 rendern.

Stage 2 — Falsifikationsversuch (openai/gpt-4o-mini, Popper-Persona)

Falsifikations-Audit

Die angegebene Falsifikationsbedingung umfasst drei symmetrische Falsifikatoren, die auf den ersten Blick eine klare Struktur bieten. Dennoch besteht die Gefahr der Immunisierung, wenn die Bedingungen nicht präzise operationalisiert werden. Insbesondere könnte die Definition der „Überlappung“ und der „Schwellen“ zu einer problematischen Flexibilität führen, die es erlaubt, die Hypothese nachträglich anzupassen, um negative Befunde zu vermeiden. Beispielsweise könnte eine Agentenpopulation, die unterhalb der 70%-Schwelle liegt, dennoch als „ausreichend überlappend“ interpretiert werden, wenn die Kriterien für die Überlappung nicht klar genug definiert sind. Zudem könnte die Verwendung von „strukturell neuen Instanzen“ als Kriterium für autonomes Lesen zu einer unklaren Operationalisierung führen, da unklar bleibt, was genau unter „strukturell neu“ zu verstehen ist. Eine präzisere Definition und Messung dieser Konzepte ist erforderlich, um die Falsifizierbarkeit der Hypothese zu gewährleisten.

Konkrete Falsifikationskandidaten

Um die Hypothese zu testen, sollten konkrete empirische Studien durchgeführt werden, die die drei genannten Konsequenzen adressieren.

1. **Funktionale Überlappung und autonome Lesbarkeit**: Eine geeignete Studie könnte Agenten mit variierenden Überlappungsgraden in einem Lesetest unterziehen, wobei die Lesefähigkeit durch standardisierte Tests wie den „Reading Comprehension Test“ gemessen wird. Ein Falsifikator wäre, wenn Agenten mit einer Überlappung unter 70% signifikant besser abschneiden als zufällig, was die Hypothese widerlegen würde.

2. **Kontrafaktische Inferenz und Autonomie-Kante**: Eine Untersuchung könnte den Zusammenhang zwischen der Fähigkeit zur kontrafaktischen Inferenz und der autonomen Lesbarkeit analysieren, indem Agenten in einem Experiment mit kontrafaktischen Szenarien konfrontiert werden. Ein Falsifikator wäre, wenn Agenten, die in der kontrafaktischen Inferenz schwach sind, dennoch in der Lage sind, neue Instanzen autonom zu lesen, was die Hypothese in Frage stellen würde.
3. **Kognitive vs. soziale Überlappungs-Schwelle**: Ein Vergleich zwischen individuellen und sozialen Leseszenarien könnte zeigen, ob die Schwellenwerte tatsächlich identisch sind. Ein Falsifikator wäre, wenn signifikante Unterschiede zwischen den Schwellen festgestellt werden, was die Substratneutralität der Hypothese widerlegen würde.

Schwellentest

Der entscheidende Schwellentest könnte wie folgt formuliert werden: Wenn Agenten mit einer funktionalen Überlappung von weniger als 70% in der Lage sind, strukturell neue Instanzen autonom zu lesen, dann ist die Hypothese widerlegt. Dies könnte durch ein Experiment geschehen, in dem zwei Gruppen von Agenten mit unterschiedlichen Überlappungsgraden (über und unter 70%) unbekannte Texte zur autonomen Lektüre erhalten. Die Lesefähigkeit könnte durch standardisierte Tests gemessen werden, und die Ergebnisse sollten statistisch verglichen werden. Ein klarer, quantitativer Befund, der zeigt, dass die Gruppe mit niedriger Überlappung signifikant besser abschneidet als erwartet, würde die Hypothese direkt in Frage stellen und ihre Falsifizierbarkeit unter Beweis stellen.

Modell: `openai/gpt-4o-mini` · Tokens: 2692 · Kosten: $0.0007 · Dauer: 14.73 s

Stage 3 — Schul-fremde Begutachtung (anthropic/claude-sonnet-4-6, Hacking-Persona)

Die vorliegende Hypothese verdient ernsthafte Aufmerksamkeit, weil sie ein echtes empirisches Problem benennt: Unter welchen Bedingungen kann ein Artefakt — eine Spur, ein Text, ein Signal — von einem Agenten gelesen werden, der den Produzenten nicht kennt und nicht anwesend hat? Das ist keine triviale Frage, und der Versuch, sie durch ein Mehr-Gruppen-Design mit vorab fixierten Falsifikatoren anzugehen, ist methodisch respektabel. Dennoch lohnt es sich, die Schul-Voraussetzungen freizulegen, die das Gerüst dieser Hypothese tragen — und die innerhalb der deutschen Anthropologie-Tradition so selbstverständlich sind, dass sie kaum als Voraussetzungen erscheinen.

Die auffälligste Vorannahme ist die Dreischichtung selbst: transzendentale Bedingung, symbolische Objektivation, funktionaler Vollzug. Diese Architektur ist nicht aus dem empirischen Material gewachsen, sondern aus einem Kanon — Kant, Cassirer, Plessner, Luhmann — der seinerseits eine bestimmte Antwort auf die Frage vorstrukturiert, was Sinnzuschreibung überhaupt ist. Wer nicht in dieser Tradition arbeitet, sieht sofort: Die Dreischichtung ist eine theoretische Entscheidung, keine empirische Entdeckung. Sie legt fest, dass es so etwas wie eine "transzendentale Bedingung" gibt, die von der sozialen Realisierung trennbar ist — und genau das ist strittig. Ein Analytiker, der mit Dennetts heterophänomenologischer Methode oder mit Churchlands eliminativem Programm arbeitet, würde fragen: Warum drei Schichten? Warum nicht eine, oder fünf? Die Antwort liegt nicht im Datenmaterial, sondern in der Schulzugehörigkeit der Autoren.

Ähnlich verhält es sich mit dem Begriff der "exzentrischen Positionalität" als Symptom-Indikator für kontrafaktische Inferenz. Plessners Konzept setzt voraus, dass es eine spezifisch menschliche Form der Selbstdistanzierung gibt, die sich von tierischer Kognition kategorial unterscheidet — eine Annahme, die in der vergleichenden Kognitionswissenschaft seit den 1990er Jahren erheblich unter Druck geraten ist. Call und Tomasello, Pepperberg, Clayton: Die Grenze zwischen kontrafaktischer Inferenz beim Menschen und bei anderen Primaten oder Corviden ist empirisch wesentlich unschärfer, als Plessners Begrifflichkeit suggeriert. Wenn der "Positionalitäts-Indikator" für Gruppe (iv) vor dem Test fixiert wird, aber auf einem kategorialen Unterschied beruht, der empirisch nicht gesichert ist, dann ist die Vorab-Fixierung kein Schutz gegen Zirkularität, sondern deren Verschleierung.

Eine konkrete Schwäche aus meiner Perspektive betrifft den Umgang mit dem, was ich — im Anschluss an Hacking — als *looping effect* bezeichnen würde: Die Klassifikation "autonomes Lesen" ist keine neutrale Beschreibung einer Eigenschaft, die Agenten unabhängig von der Klassifikation besitzen. Sobald Agenten wissen, dass sie auf "autonomes Lesen" getestet werden, verändern sie ihr Verhalten — und das gilt für menschliche Versuchspersonen ebenso wie, auf andere Weise, für trainierte KI-Systeme, deren Trainingskorpora die Klassifikationspraxis der Forscher bereits enthalten. Die Hypothese behandelt "Rezeptionsmarker" und "Produktionsmarker" als ob sie stabile Messgegenstände wären; sie sind es nicht. Sie sind Ko-Konstruktionen aus Messinstrument und Messgegenstand. Das ist kein philosophisches

Glasperlenspiel: Es hat direkte Konsequenzen für das Mehr-Gruppen-Design, weil die Gruppen nicht dieselbe Klassifikation stabil realisieren, wenn die Klassifikation selbst durch den Test miterzeugt wird.

Die produktive Anschlussidee, die ich aus meiner eigenen Tradition anbieten möchte, ist die der *historischen Ontologie* im Sinne Hackings: Statt zu fragen, ob "autonomes Lesen" eine natürliche Art ist, die durch Überlappungstiefe realisiert wird, sollte man fragen, wann und wie "autonomes Lesen" als Kategorie in wissenschaftlichen und sozialen Praktiken auftauchte, welche Institutionen sie stabilisierten, und welche Objekte sie dadurch erst erzeugten. Der Schrift-Übergang, den die Hypothese als "historischen Skalierungstest" verwendet, ist in dieser Lesart kein Bestätigungsfall, sondern ein Entstehungsfall: Schrift hat nicht eine vorher existierende Fähigkeit zum autonomen Lesen skaliert, sondern eine neue Art von Agenten — den Leser — erst hervorgebracht, zusammen mit neuen Institutionen (Schule, Bibliothek, Kommentar), die diese Art stabilisierten. Das ist empirisch prüfbar, aber mit anderen Methoden: nicht durch Schwellenmessung, sondern durch historisch-epistemologische Rekonstruktion der Praktiken, in denen "Lesen" als Kategorie operativ wurde. Dieser Zugang würde die Substratneutralitätsthese nicht widerlegen, aber erheblich präzisieren — und den Reihenfolge-Streit zwischen "Modell vor Anschluss" und "Modell aus dem Anschluss" durch eine dritte Option ersetzen: Modell und Anschluss entstehen gemeinsam in einer Praxis, die historisch datierbar und institutionell verankert ist.

Modell: `anthropic/claude-sonnet-4-6` · Tokens: 3330 · Kosten: $0.0285 · Dauer: 32.2 s

Wirkung auf die finale Bewertung

Doppelausweisung intern / nach externer Prüfung im Hauptbericht (Frontmatter `finale_summe` und `finale_summe_nach_externer_pruefung`). Korrekturen werden im Anhang explizit begründet, nicht im internen Bewertungsblock überschrieben. Eine inhaltliche Auswertung der Stage-Outputs erfolgt durch den User; gegebenenfalls werden Befunde als zusätzliche Reservoir-Verzweigungen mit `#verzweigung-offen-extern-<thema>` und `quelle: extern-stage<n>` dokumentiert.

Korrekturbegründung der finalen Bewertung

Intern: 76/90. Nach externer Prüfung: 74/90 (−2).

Stage 1 (Originalität, perplexity/sonar-reasoning-pro) fiel mit einem Upstream-502 (Perplexity timed out) aus; Skip-Marker gesetzt, Stage 2/3 gerendert. Die Originalitätsdimension wurde daher extern nicht geprüft — der interne Wert bleibt unkorrigiert stehen.

Die Korrektur um −2 stützt sich auf zwei substanzielle Befunde: - **Stage 3 (Hacking-Persona)** zeigt einen looping effect auf die Messgröße selbst: „autonomes Lesen" ist keine vom Klassifikationsakt unabhängige Eigenschaft, Rezeptions- und Produktionsmarker sind Ko-Konstruktionen aus Messinstrument und Messgegenstand. Das senkt **Begriffliche Klarheit** (8 → 7). Zudem steht der vorab fixierte Positionalitäts-Indikator für Gruppe (iv) unter empirischem Druck (kategoriale Mensch/Tier-Grenze unscharf; Tomasello, Clayton) — die Vorab-Fixierung schützt dann nicht gegen Zirkularität, sondern verschleiert sie. Das senkt **Falsifizierbarkeit** (9 → 8). - **Stage 2 (Popper-Persona)** bestätigt die Falsifikationsstruktur als methodisch respektabel, mahnt aber dieselbe Operationalisierungs-Flexibilität an (Schwellen-/„strukturell neu"-Definitionen).

Neue Summe: 73 (intern Auswahl) → 76 (nach Runden) → **74 nach externer Prüfung**. Die Korrektur überschreibt nicht den internen Bewertungsblock; die Doppelausweisung steht im Frontmatter (`finale_summe` 76, `finale_summe_nach_externer_pruefung` 74).

Vollständige Stage-Outputs: `state/zweitmeinung-2026-06-21.md`.