Hypothesentag 2026-06-03 — Der Ort der Normativität

1. Verzweigungs-Pickup

warm_pick. `empirie-pr` (erfasst 2026-05-20, Makro-Thema Erkenntnistheorie_Cassirer) — aus 06 Hypothesentag/2026-05-20.

cold_pick (Diversitäts-Pflicht). `symbolische-funktion-kunst-differenz` (erfasst 2026-05-25, Makro-Thema Ästhetik_Goethe) — Makro-Rotation: 'Ästhetik_Goethe' war in den letzten 5 Tagen nicht vertreten.

devil_advocate_pick (Mittwochs-Außenblick). Die regulationstheoretische Achse begeht den Fehler, den Wittgensteins Regelfolge-Paradox diagnostiziert: Sie reifiziert die Norm der Korrektheit zu einem verborgenen Mechanismus (Sollwert, symbolische Funktion), obwohl keine interne Struktur die Extension einer Regel für neue Fälle festlegen kann. Korrektheit ist kein im Organismus installierter Sollwert, sondern ein Merkmal davon, wie eine Praxisgemeinschaft fortfährt — sie liegt in der Lebensform, nicht im generativen Modell.

- Wochenschwerpunkt: Die Woche wurde dominiert von der epistemologisch-systemtheoretischen Klassifikation, mit Fokus auf Cassirer - Regelung als transzendentales Prinzip und Powers - Höhlenmalerei als Sollwertinstallation, wiederkehrend das Vokabular Sollwert, Konstitutionslogik, aktive Inferenz und 'symbolische Funktion'.
- Paradigmatische Kernannahme: Stillschweigend vorausgesetzt wurde, dass die Normativität von Kultur und Psyche — was als adäquate Realisierung, korrekte Fortsetzung oder gelungene Symbolisierung gilt — durch eine interne, regulative Struktur festgelegt ist: einen Sollwert, ein generatives Modell oder eine 'symbolische Funktion', die den Cases vorausliegt und bestimmt, welche von ihnen sie korrekt erfüllen.
- Alternative Tradition: Wittgensteins Spätphilosophie (Regelfolgen, Lebensform, Privatsprachenargument) gegen die Cassirer-Friston-Achse
- Aufhänge-Knoten: Cassirer - Regelung als transzendentales Prinzip (sowie der cold-pick-Knoten 'symbolische Funktion' aus 2026-05-25)

2. Drift-Warnung und Echo-Chamber-Alarm

Drift-Warnung: keine. Echo-Chamber-Alarm: keiner. Devil's Advocate dennoch aktiv (planmäßiger Mittwochs-Außenblick). Schwerpunkt der Woche: epistemologisch_systemtheoretisch.

3. Drei Hypothesen (initial)

Hypothese 1 (warm_pick: empirie-pr) — Der phänomenale Rest als Konstitutionsbedingung, nicht als Residuum

Kernsatz. Der "phänomenale Überschuss" der primären Intersubjektivität, den das Cassirer-Friston-Programm bisher als ungeklärtes Residuum behandelt, ist kein Rest, der nach vollständiger Modellierung übrigbleibt, sondern die Konstitutionsbedingung dafür, dass aktive Inferenz überhaupt auf *bedeutsame* Welt gerichtet sein kann: Prädiktives Kodieren optimiert Priors innerhalb einer bereits affektiv erschlossenen Welt, kann aber die Erschlossenheit selbst nicht aus sich erzeugen.

Begründung. Die Verzweigung `empirie-pr` (aus 06 Hypothesentag/2026-05-20) hält die Streitfrage offen, ob ein vollständig formalisiertes Active-Inference-Modell die synchrone Affektivität des Säuglings (Trevarthen) ohne phänomenologischen Rest rekonstruieren kann. Die übliche Lesart behandelt den Rest als Mess- oder Modellierungslücke, die sich prinzipiell schliessen liesse. Die Hypothese kehrt die Beweislast um: Sie knüpft an Cassirer - Regelung als transzendentales Prinzip und die dort markierte Konstitutionslücke an und behauptet, dass die affektiv-leibliche Erschlossenheit logisch *vor* der Prior-Optimierung liegt — sie ist das, worauf sich Inferenz richtet, nicht ihr Produkt. Damit wird der Rest vom empirischen Defizit zum transzendentalen Argument.

Falsifikationsbedingung. Widerlegt, wenn eine Friston-Rekonstruktion von Trevarthen-Protokollen die interleibliche Kalibrierung des Säuglings vollständig aus interozeptivem predictive coding ableitet, *ohne* eine vorgängige Welterschlossenheit als nicht-abgeleitete Randbedingung des generativen Modells voraussetzen zu müssen.

Quelle. warm_pick — verfolgt den Strang aus 06 Hypothesentag/2026-05-20.

Hypothese 2 (cold_pick: symbolische-funktion-kunst-differenz) — Das Urphänomen als nicht-funktionaler Grund der Wirkungsstruktur

Kernsatz. Goethes Urphänomen ist keine symbolische Funktion und lässt sich auch nicht als solche rekonstruieren: Es bezeichnet einen Punkt, an dem die Wirkungsstruktur eines Phänomens nicht mehr funktional — also nicht mehr durch eine Regel, die festlegt, welche Realisierungen als adäquat gelten — bestimmt ist, sondern anschaulich-unmittelbar gegeben. Die Wirkungsontologie des Kunstwerks braucht deshalb eine Schicht *unterhalb* der symbolischen Funktion.

Begründung. Die Verzweigung `symbolische-funktion-kunst-differenz` (aus 06 Hypothesentag/2026-05-25) folgt der These, das Kunstwerk sei eine symbolische Funktion im Sinne Cassirers, die bestimmt, welche Wirkungen es korrekt realisieren. Genau diese Bestimmungsrelation greift die Gegenbewegung an: Bei Goethe ist das Urphänomen das, "wo nichts weiter dahinter gesucht werden soll" — ein anschaulicher Grenzbegriff, an dem das

Bestimmen aufhört und das Schauen beginnt. Überträgt man das auf die Kunst, dann gibt es in jedem Werk eine Schicht, die nicht festlegt, welche Rezeption adäquat ist, sondern die als sinnlich-sittliche Polarität (Farbenlehre) unmittelbar wirkt. Die Hypothese stellt damit gegen die funktionalistische Lesung der Ästhetik-Woche eine goetheanische: Die Identität des Werks gründet nicht durchgängig in einer regelhaften Funktion, sondern endet in einem anschaulichen Urphänomen.

Falsifikationsbedingung. Widerlegt, wenn sich die als "urphänomenal-unmittelbar" beanspruchte Wirkungsschicht (etwa die affektive Sofortwirkung von Farbkontrasten) restlos als implizite Regel rekonstruieren lässt, die festlegt, welche Rezeptionen das Werk korrekt realisieren — dann wäre auch sie symbolische Funktion und keine eigene Schicht.

Quelle. cold_pick — Diversitäts-Pflicht. Makro-Thema Ästhetik_Goethe, in den letzten 7 Tagen nicht berührt.

Hypothese 3 (devil_advocate) — Normativität liegt in der Lebensform, nicht im Sollwert

Gegenthese-zu. epistemologisch-systemtheoretischer Wochenschwerpunkt (Cassirer-Friston-Powers-Achse, "Sollwert" / "symbolische Funktion").

Kernsatz. Die regulationstheoretische Achse begeht den Fehler, den Wittgensteins Regelfolge-Paradox diagnostiziert: Sie reifiziert die Norm der Korrektheit zu einem internen Mechanismus — Sollwert, generatives Modell, symbolische Funktion —, obwohl keine interne Struktur die Extension einer Regel für genuin neue Fälle festlegen kann. Korrektheit ist kein im Organismus installierter Sollwert, sondern ein Merkmal davon, wie eine Praxisgemeinschaft fortfährt; sie liegt in der Lebensform, nicht im Regelkreis.

Begründung. Der Wochenschwerpunkt setzt durchgängig voraus, dass das, was als adäquate Realisierung, korrekte Fortsetzung oder gelungene Symbolisierung gilt, durch eine den Fällen vorausliegende interne Struktur bestimmt sei — so Cassirer - Regelung als transzendentales Prinzip und die "symbolische Funktion" aus 06 Hypothesentag/2026-05-25. Wittgensteins Regelfolge-Betrachtungen (PU §§185–242) zeigen aber: Jede endliche interne Spezifikation lässt sich auf unendlich viele Weisen fortsetzen; was die Fortsetzung als korrekt auszeichnet, kann nicht noch einmal eine innere Repräsentation sein, ohne den Regress zu eröffnen. Der "Sollwert" erklärt die Korrektheit nur, wenn man schon weiss, wie er anzuwenden ist — und dieses Wissen ist Praxis, nicht Mechanismus. Die Gegenthese verlagert den Ort der Normativität von der Friston-Architektur in die geteilte Lebensform.

Falsifikationsbedingung. Widerlegt, wenn sich für eine genuin neue kulturelle Normanwendung zeigen lässt, dass ein vollständig spezifiziertes generatives Modell (Sollwert) die korrekte Fortsetzung vorab eindeutig fixiert — ohne jeden Rekurs auf

gemeinschaftliche Aufnahme (uptake). Dann läge die Korrektheit doch in der internen Struktur.

Quelle. devil_advocate — Mittwochs-Außenblick gegen den Wochenschwerpunkt.

4. Kritischer Professor pro Hypothese

Drei Vorwurfs-Sets. Methodik: Falsifikationsprüfung, alternative Erklärung, Begriffsschärfe, Zirkularitätskontrolle. Keine normative Bewertung.

H1 — Phänomenaler Rest als Konstitutionsbedingung

Vorwurf 1 — Transzendentaler Trick statt Falsifizierbarkeit. Die These macht die Erschlossenheit zur "Konstitutionsbedingung", die der Inferenz logisch vorausliegt. Damit ist sie gegen jeden empirischen Befund immunisiert: Gelingt eine Friston-Rekonstruktion doch, lässt sich immer behaupten, die Erschlossenheit sei als stillschweigende Randbedingung schon eingebaut. Die Falsifikationsbedingung versucht das zu kontern ("ohne vorgängige Welterschlossenheit als nicht-abgeleitete Randbedingung"), aber wer entscheidet, ob eine Randbedingung "nicht-abgeleitet" ist? Ohne Operationalisierung dieses Kriteriums ist die These ein Antinomie-Argument, keine empirische Hypothese.

Vorwurf 2 — Äquivokation bei "Rest". "Phänomenaler Überschuss" changiert zwischen (a) erklärungsresistentem Qualia-Rest und (b) logischer Vorbedingung. Das sind zwei verschiedene Behauptungen. (a) ist die harte Bewusstseinslücke (Chalmers), (b) ein transzendentales Argument (Cassirer). Die These verkauft (b) und lehnt sich an die Plausibilität von (a) an.

Vorwurf 3 — Alternative Erklärung. Enaktivismus (Di Paolo, Thompson) erklärt Erschlossenheit als Resultat sensomotorischer Kopplung ohne phänomenalen Sonderstatus. Die These muss zeigen, warum Erschlossenheit *nicht* dynamisch-relational rekonstruierbar ist — sonst ist der "Überschuss" nur ein noch nicht modelliertes Kopplungsphänomen.

H2 — Urphänomen als nicht-funktionaler Grund

Vorwurf 1 — Begriffliche Gleichsetzung Goethe/Kunst nicht ausgewiesen. Das Urphänomen ist bei Goethe ein naturwissenschaftlicher Grenzbegriff (Farbenlehre, Morphologie). Die Übertragung auf das Kunstwerk ist eine Analogie, kein Argument. Warum sollte ein Werk eine "urphänomenale Schicht" haben, nur weil Naturphänomene einen anschaulichen Grund haben? Der Schritt von der Naturanschauung zur Kunstontologie ist die eigentliche Beweislast und bleibt offen.

Vorwurf 2 — Selbstuntergrabung durch die Falsifikationsbedingung. Die These räumt selbst ein: Wenn die "unmittelbare" Wirkungsschicht restlos als implizite Regel rekonstruierbar ist, ist sie symbolische Funktion. Die Wahrnehmungspsychologie der

Farbwirkung (Kontrastregeln, opponente Prozesse) legt aber genau das nahe — Farbsofortwirkung ist hochgradig gesetzmässig. Die These steht damit empirisch auf dünnem Eis und droht in die funktionalistische Lesart zurückzufallen, die sie bekämpft.

Vorwurf 3 — "Anschaulich-unmittelbar" ist kein Prädikat mit Kontrast. Solange nicht gesagt wird, was es hiesse, dass eine Wirkung *nicht* unmittelbar ist, bleibt der Kernbegriff leer. Mittelbarkeit/Unmittelbarkeit braucht ein operationalisierbares Kriterium (z. B. Latenz, Kontextabhängigkeit, Lernbarkeit), sonst ist die behauptete "Schicht unterhalb der Funktion" nicht identifizierbar.

H3 — Normativität in der Lebensform (Devil's Advocate)

Vorwurf 1 — Kripkes Skeptizismus trifft die Lebensform genauso. Wenn keine endliche interne Struktur die Extension fixiert, dann fixiert sie auch keine endliche Praxis: Auch das bisherige Gemeinschaftsverhalten ist endlich und lässt unendlich viele Fortsetzungen zu (Kripkensteins Quaddition gilt für "die Gemeinschaft geht so fort" ebenso). Die These verlagert das Problem nur, statt es zu lösen. Sie muss zeigen, warum die Lebensform der Unterbestimmtheit entgeht, der der Sollwert erliegt.

Vorwurf 2 — Strohmann-Gefahr. Friston-Theoretiker behaupten nicht, der Sollwert lege die Extension "vorab eindeutig" fest; aktive Inferenz ist probabilistisch und kontextsensitiv, das generative Modell wird *durch* Interaktion mit der Umwelt (auch der sozialen) geformt. Die Gegenthese unterstellt einen statischen, internalistischen Sollwert-Begriff, den die avancierte Theorie gar nicht vertritt. Damit zielt die Falsifikationsbedingung ("ohne jeden Rekurs auf uptake") auf eine Position, die niemand hält.

Vorwurf 3 — Falsche Dichotomie intern/sozial. "Lebensform" und "generatives Modell" sind keine Alternativen, wenn das generative Modell soziale Priors enthält (vgl. den Vault-Knoten zur interaktiven Gain-Modulation sozialer Prioren). Dann ist die geteilte Praxis *im* Modell repräsentiert, und der behauptete Gegensatz kollabiert. Die These muss einen Punkt benennen, an dem soziale Praxis nicht in Priors überführbar ist — sonst ist sie mit der Gegenposition verträglich statt ihr entgegengesetzt.

5. Reformulierung

Jede Hypothese überarbeitet: Schwachstellen behoben durch Schärfung, nicht durch Verschleierung. Quellen-Tags bleiben erhalten.

H1 (warm_pick: empirie-pr) — Die Erschlossenheits-Asymmetrie als operationalisierbares Forschungsproblem

Kernsatz. Aktive Inferenz erklärt, wie ein Organismus Priors über eine bedeutsame Welt optimiert, aber sie setzt die Bedeutsamkeit der Welt als nicht-abgeleitete Randbedingung des

generativen Modells voraus; diese Asymmetrie ist nicht ein metaphysischer "Überschuss", sondern ein präzises Defizit: Kein Active-Inference-Modell der primären Intersubjektivität kommt ohne eine bereits affektiv gewichtete Anfangsverteilung aus, deren Gewichtung es selbst nicht generiert.

Begründung. (Antwort auf Vorwurf 1 + 2.) Die These verzichtet auf den Qualia-Begriff und auf das transzendentale "logisch-vorausliegend". Sie behauptet stattdessen ein modelltechnisches Faktum: Jedes generative Modell braucht eine Prior-Initialisierung; bei der Säuglings-Intersubjektivität (Trevarthen) ist diese Initialisierung affektiv strukturiert (selektive Reagibilität auf Gesichter, Stimmprosodie, Kontingenz). Die offene Frage `empirie-pr` aus 06 Hypothesentag/2026-05-20 und die Konstitutionslücke in Cassirer - Regelung als transzendentales Prinzip werden damit zu einer empirischen Frage über die Herkunft der Prior-Gewichtung. (Antwort auf Vorwurf 3:) Der Enaktivismus ist nicht Gegner, sondern Testfall — er behauptet, die Gewichtung entstehe aus sensomotorischer Kopplung; die These ist falsch, falls er recht hat.

Falsifikationsbedingung. Operationalisiert: Widerlegt, wenn ein Active-Inference- oder enaktives Modell die affektive Anfangsgewichtung der Säuglings-Priors aus voraffektiven sensomotorischen Kontingenzen ableitet, ohne sie als Parameter vorab zu setzen — messbar daran, dass das Modell die selektive Gesichts-/Prosodie-Reagibilität Neugeborener ohne eingebauten affektiven Bias reproduziert.

Quelle. warm_pick — verfolgt den Strang aus 06 Hypothesentag/2026-05-20.

H2 (cold_pick: symbolische-funktion-kunst-differenz) — Latenz und Lernbarkeit als Trennkriterium zweier Wirkungsschichten

Kernsatz. In der Wirkungsstruktur des Kunstwerks lassen sich zwei Schichten durch ein operationalisierbares Kriterium trennen: eine *präreflexive* Schicht (kurze Latenz, geringe Kontextabhängigkeit, kaum lernabhängig — analog zu Goethes anschaulichem Urphänomen) und eine *symbolisch-funktionale* Schicht (kontextabhängig, lernbar, regelhaft). Die These behauptet nicht, die erste Schicht sei regelfrei, sondern dass sie nicht durch eine *werk-konstitutive* Norm bestimmt ist, die festlegt, welche Rezeption das Werk korrekt realisiert.

Begründung. (Antwort auf Vorwurf 1:) Die Goethe-Analogie wird zum begrenzten heuristischen Modell zurückgestuft — das Urphänomen liefert die Idee einer Schicht, "wo nichts dahinter zu suchen ist", nicht den Beweis. (Antwort auf Vorwurf 2 + 3:) Der Einwand der Wahrnehmungspsychologie wird übernommen, nicht abgewehrt: Ja, Farbsofortwirkung ist gesetzmässig (opponente Prozesse). Aber gesetzmässig ≠ werk-konstitutiv. Eine spezieswe perzeptuelle Regel legt nicht fest, welche Rezeption *dieses Werk* korrekt realisiert; das tut erst die symbolische Funktion (06 Hypothesentag/2026-05-25). Das Trennkriterium ist damit Latenz/Kontextsensitivität/Lernbarkeit — drei messbare Grössen.

Falsifikationsbedingung. Widerlegt, wenn sich keine empirisch trennscharfe Schicht-Grenze findet — d. h. wenn die als präreflexiv beanspruchten Wirkungen denselben Grad an Kontextabhängigkeit und Lernbarkeit zeigen wie die symbolisch-funktionalen, sodass die Zweischichtung perzeptuell nicht messbar ist.

Quelle. cold_pick — Diversitäts-Pflicht. Makro-Thema Ästhetik_Goethe.

H3 (devil_advocate) — Die Verkörperungs-Differenz: Praxis bindet, weil sie sanktioniert, nicht weil sie repräsentiert

Gegenthese-zu. epistemologisch-systemtheoretischer Wochenschwerpunkt.

Kernsatz. Geteilte Praxis löst die Regelfolge-Unterbestimmtheit nicht dadurch, dass sie die korrekte Fortsetzung *repräsentiert* (das könnte ein generatives Modell mit sozialen Priors auch), sondern dadurch, dass sie *sanktioniert*: Korrektheit ist konstituiert durch reale Korrekturen, Ausschlüsse und Bestätigungen einer Gemeinschaft, die kein Modell im Kopf eines Einzelnen ersetzen kann, weil sie ein Ereignis zwischen Körpern ist. Der Sollwert kann die Sanktion modellieren, aber nicht *sein*.

Begründung. (Antwort auf Vorwurf 1 — Kripke:) Die These beansprucht nicht mehr, die Lebensform "fixiere die Extension"; das kann sie nicht, und das ist der Punkt. Korrektheit ist nichts Fixiertes, sondern etwas fortlaufend Durchgesetztes. Der Skeptizismus trifft nur Theorien, die eine Determination *suchen* — die Sanktions-These sucht keine. (Antwort auf Vorwurf 2 + 3 — Strohmann/falsche Dichotomie:) Zugestanden, das generative Modell enthält soziale Priors (vgl. Reservoir - Interaktive Gain-Modulation soziale Prioren 2026-05-27). Aber ein Prior *über* Sanktionen ist nicht die Sanktion. Die Differenz ist konkret: Ein perfektes internes Modell einer Praxisgemeinschaft, das in einem isolierten System läuft, hätte normative Korrektheits-*Erwartungen*, aber keine Korrektheit — denn niemand könnte es korrigieren. Das ist der Punkt, an dem soziale Praxis nicht in Priors überführbar ist: Die Überführung verliert die zweite Person.

Falsifikationsbedingung. Widerlegt, wenn ein einzelnes, sozial isoliertes System gezeigt werden kann, das den vollen normativen Status (richtig/falsch, nicht nur erwartet/unerwartet) seiner Begriffe allein aus internen Priors erzeugt — ohne dass eine zweite Instanz seine Anwendungen je korrigieren könnte.

Quelle. devil_advocate — Mittwochs-Außenblick.

6. Bewertung mit dreifacher Klassifikation

H1 — Erschlossenheits-Asymmetrie

Themenfeld: epistemologisch_systemtheoretisch_hypothese (Prior-Architektur, generatives Modell).
Methoden-Typ: empirisch_pruefbar — behauptet ein modelltechnisches Faktum über die Herkunft der Prior-Gewichtung; Variablen (Gesichts-/Prosodie-Reagibilität) messbar.
Pflichttest Operationalisierung: bestanden — affektive Anfangsgewichtung über Neugeborenen-Reagibilität messbar.
Reichweiten-Klasse: these — erweitert 2 Vault-Knoten, klare Falsifikation, eigenständig publizierbar.

Kriterium	Score	Begründung
Originalität	7	Reframing einer bekannten PC-Debatte (Konstitutionslücke), nicht völlig neu, aber zugespitzt zur Prior-Herkunftsfrage
Falsifizierbarkeit	9	Klarer Schwellentest: Ableitung der affektiven Gewichtung ohne eingebauten Bias
Begriffliche Klarheit	8	"Anfangsgewichtung" und "nicht-abgeleitete Randbedingung" sauber, Qualia-Begriff bewusst vermieden
Tiefe	8	Trifft das transzendentale Fundament der aktiven Inferenz
Forschungsrelevanz	9	Direkt anschlussfähig an Friston/Hohwy/enaktivistische Debatte 2022+
Interdisziplinär	8	Entwicklungspsychologie, Kognitionswissenschaft, Phänomenologie
Vault-Anschluss	9	Knüpft eng an Cassirer-Knoten und empirie-pr-Verzweigung
Antinomie-Test	7	Enaktivismus liefert ernste Gegenposition
Publikationsmöglichkeit	8	Solide, aber inkrementeller Beitrag
Summe	**73**	

H2 — Latenz/Lernbarkeit als Schicht-Trennkriterium

Themenfeld: asthetisch_anschauliche_hypothese (Wirkungsstruktur des Werks).
Methoden-Typ: empirisch_pruefbar — Schicht-Trennung über Latenz/Kontextsensitivität/Lernbarkeit messbar; ästhetische Rahmung.
Pflichttest Operationalisierung: bestanden — drei messbare Grössen benannt.
Reichweiten-Klasse: these — erweitert den Knoten symbolische Funktion um eine messbare Untergrenze.

Kriterium	Score	Begründung
Originalität	8	Goethe-Urphänomen als perzeptuelle Untergrenze der symbolischen Funktion ist ein frischer Schnitt
Falsifizierbarkeit	7	Trennschärfe-Test plausibel, aber Operationalisierung von "werk-konstitutiv" bleibt anspruchsvoll
Begriffliche Klarheit	7	"werk-konstitutiv vs. speziesweit" geklärt, aber Restunschärfe in "präreflexiv"
Tiefe	7	Greift in Kunstontologie, bleibt aber unterhalb der Grundfrage
Forschungsrelevanz	7	Anschluss an empirische Ästhetik (Provenienzeffekt), mittlere Aktualität
Interdisziplinär	8	Philosophie + Wahrnehmungspsychologie + Kunstgeschichte
Vault-Anschluss	7	Greift cold_pick-Knoten auf, aber aus der Wochenschneise entfernt
Antinomie-Test	7	Funktionalistische Gegenlesart stark
Publikationsmöglichkeit	7	Publizierbar, aber Spezialdebatte
Summe	**65**	

H3 — Sanktion vs. Repräsentation (Devil's Advocate)

Gegenthese-zu: epistemologisch-systemtheoretischer Wochenschwerpunkt.
Themenfeld: epistemologisch_systemtheoretisch_hypothese (Lokalisierung von Normativität in Regel-/Modellarchitektur).
Methoden-Typ: begriffsanalytisch — dominanter Prüfweg ist die Analyse von "Korrektheit", "Repräsentation", "Sanktion"; empirische Randkomponente (isoliertes System) vermerkt.
Pflichttest Definitions-Schärfe: bestanden — Unterscheidung repräsentieren/sanktionieren trennscharf, zweite-Person-Argument als Scharnier.
Adjustierte Anker: Falsifizierbarkeit max 7, Begriffliche Klarheit max 10, Antinomie-Test max 10.
Reichweiten-Klasse: these — eine zugespitzte Einzelaussage mit klarer Falsifikation; öffnet aber die Lokalisierungs-Frage breit (Forschungsprogramm-Potential vermerkt).

Kriterium	Score	Begründung
Originalität	9	Wittgenstein-Sanktionsbegriff scharf gegen die Friston-Sollwert-Achse gestellt — direkter Paradigmen-Konflikt

Kriterium	Score	Begründung
Falsifizierbarkeit	6	begriffsanalytisch gedeckelt (max 7); Test über isoliertes System tragfähig, aber idealisiert
Begriffliche Klarheit	9	repräsentieren/sanktionieren und erwartet/korrekt sauber getrennt
Tiefe	9	Trifft die Grundannahme der ganzen Wochenachse: Ort der Normativität
Forschungsrelevanz	8	Lebendige Debatte (normativity of predictive processing, Bruineberg, Constant)
Interdisziplinär	8	Sprachphilosophie, Kognitionswissenschaft, Sozialontologie
Vault-Anschluss	7	Hängt an Cassirer-Knoten und soziale-Prioren-Reservoir, ist aber dezidiert kontra
Antinomie-Test	9	Die Gegenposition (Priors über Sanktionen) ist stark — gerade das macht den Test wertvoll
Publikationsmöglichkeit	8	Pointierte Kontra-These, gut platzierbar
Summe	**73**	

Auswahl

Gleichstand H1 und H3 bei **73**. Tiebreak laut Regel: höhere **Originalität** entscheidet → H3 (9) vor H1 (7).

Gewählte Hypothese: H3 — Sanktion vs. Repräsentation (Devil's Advocate). Reichweiten-Klasse der gewählten Hypothese: these.

Reservoir-Profile (nicht gewählt)

- **H1 — Erschlossenheits-Asymmetrie** (Score 73, these): hochwertiger Strang, direkt wieder aufgreifbar. Kein forschungsprogramm_kandidat, aber Pickup-Priorität hoch.
- **H2 — Schicht-Trennkriterium** (Score 65, these): Ästhetik-Strang, anschlussfähig an empirische Ästhetik. forschungsprogramm_kandidat: false.

7. Expertenrunde 1 — unabhängige Gutachten

Panel-Konfiguration. Hypothese H3 verhandelt die Lokalisierung von Normativität in einer Regel- bzw. Modellarchitektur — epistemologisch-systemtheoretisches Standardpanel. Wittgenstein ist Stichwortgeber, Friston und Cassirer die unmittelbar Betroffenen.

Panel-Mitglieder. Kant, Popper, Wittgenstein, Cassirer, Friston, Luhmann, Goethe.

Kant

Die These trifft eine echte Bedingung der Möglichkeit, verfehlt aber deren Ort. Korrektheit ist in der Tat nicht aus einer endlichen Vorstellung ableitbar — das ist nichts anderes als das alte Problem der Subsumtion: keine Regel kann die Regel ihrer eigenen Anwendung mitliefern, sonst droht der Regress, den ich der Urteilskraft als eigenem Vermögen zugewiesen habe. Insofern ist die Diagnose richtig. Doch der Schluss, die Korrektheit liege deshalb in der äußeren Sanktion der Gemeinschaft, vertauscht das Konstitutive mit dem Empirischen. Sanktion ist Faktum, nicht Geltung. Daß eine Gemeinschaft jemanden korrigiert, erklärt nicht, mit welchem Recht sie es tut — sonst wäre jede durchgesetzte Übereinkunft schon richtig, und der Unterschied zwischen 'man hält es für richtig' und 'es ist richtig' verschwände. Mein Einwand ist also präzise: Die These braucht eine Instanz, die der Sanktion ihre Verbindlichkeit gibt, und diese Instanz ist nicht selbst wieder eine Sanktion. Schärfen ließe sich die These, wenn sie zwischen der Genese der Korrektheits-Praxis und ihrem Geltungsgrund unterschiede. Die zweite Person, die hier so stark gemacht wird, ist verwandt mit dem, was ich Achtung nenne — aber Achtung ist eine Anerkennung des Gesetzes, nicht der Person. Untersuchen Sie, ob nicht die Sanktion bloß die sinnliche Erscheinung einer überpersonalen Verbindlichkeit ist.

Popper

Mir gefällt, daß hier überhaupt eine Falsifikationsbedingung steht — das isolierte System, das vollen normativen Status aus internen Priors erzeugte, wäre tatsächlich ein Widerleger. Aber ich bezweifle, daß diese Bedingung je eintreten könnte, und das ist verdächtig. Eine These, deren Falsifikator prinzipiell unrealisierbar ist, nähert sich der Unwiderlegbarkeit. Was hieße 'voller normativer Status, ohne daß jemand korrigieren könnte'? Wer entscheidet, daß das System richtig und nicht nur konsistent operiert? Hier droht Immunisierung durch Definition: Sie haben Korrektheit bereits an Korrigierbarkeit gekoppelt, also kann kein isoliertes System sie per definitionem erreichen. Das macht die Falsifikation zirkulär. Mein konstruktiver Vorschlag: Ersetzen Sie das Gedankenexperiment durch eine prüfbare Asymmetrie. Etwa — Lerner, die nie sozial korrigiert werden, sollten systematische, nicht zufällige Normabweichungen entwickeln, die intern nicht erkennbar sind. Das ist messbar, etwa bei sprachlich isolierten oder rein selbstüberwachten Lernsystemen. Übrigens ist Ihre Position der meinen näher als Sie denken: Kritik ist intersubjektiv, Erkenntnis wächst durch Widerlegung von außen. Aber ich würde nie sagen, die Wahrheit liege in der kritisierenden Gemeinschaft — sie liegt jenseits ihrer, und die Gemeinschaft ist nur das Verfahren, ihr näherzukommen. Trennen Sie das Verfahren vom Status, dann steht die These.

Wittgenstein

Man liest hier den richtigen Satz und zieht aus ihm eine Theorie — und genau das wollte ich nicht. 'Korrektheit liegt in der Lebensform' ist keine Behauptung über einen Ort, an dem

etwas liegt. Es ist eine Erinnerung daran, daß die Frage 'wo liegt sie?' selbst schon entgleist ist. Die These macht aus der Sanktion ein neues Versteck für das, was man im Sollwert suchte. Aber 'sanktionieren' ist nicht fundamentaler als 'der Regel folgen'; es ist eine der vielen Handlungen, in denen sich zeigt, was wir eine Regel nennen. Der gefährliche Schritt ist 'konstituiert durch'. Nichts wird hier konstituiert. Wir folgen der Regel blind — und das Korrigiertwerden gehört zu diesem Spiel, ist aber nicht sein Grund. Was mir gefällt: die zweite Person, das Korrigieren als Szene zwischen Menschen, nicht im Kopf. Das ist gegen den Mentalismus gut gesagt. Mein Einwand ist nur: Sagen Sie nicht, das sei der Mechanismus der Normativität. Es gibt keinen Mechanismus. Schärfen ließe sich das, indem Sie 'konstituiert durch Sanktion' streichen und durch 'zeigt sich im Sanktionieren' ersetzen. Dann ist es keine Konkurrenztheorie zu Friston mehr, sondern die Feststellung, daß seine Theorie und meine über verschiedene Dinge reden: Er beschreibt einen Apparat, ich eine Praxis. Lesen Sie §201 noch einmal — es gibt eine Auffassung der Regel, die nicht Deutung ist.

Cassirer

Ich erkenne meinen Begriff der symbolischen Funktion als Angeklagten wieder und möchte ihn verteidigen, aber nicht, indem ich ihn internalisiere — das tut die These mir zu Unrecht. Die symbolische Funktion war nie ein Mechanismus im Kopf eines Einzelnen. Sie ist eine Form der Synthesis, die zwischen Sinnlichem und Bedeutung vermittelt, und sie ist von Anfang an in einer Kulturgemeinschaft objektiviert — in Sprache, Mythos, Kunst. Insofern ist die Entgegensetzung 'Lebensform statt symbolische Funktion' eine Scheinopposition: Die Lebensform ist der Ort, an dem symbolische Formen leben. Was die These richtig sieht, ist die Unhintergehbarkeit der Gemeinschaft. Was sie übersieht, ist, daß die Gemeinschaft selbst nur durch geteilte symbolische Formen Gemeinschaft ist — Sanktion setzt schon eine geteilte Bedeutung dessen voraus, was sanktioniert wird. Der Einwand ist also: Die Sanktion ist nicht der Grund der Korrektheit, sondern bereits ein symbolischer Akt, der eine Norm voraussetzt, statt sie zu stiften. Sonst sanktionierte man nur Verhalten, nicht Falschheit. Mein Vorschlag zur Schärfung: Fragen Sie, wie eine Sanktion überhaupt als Korrektur eines Fehlers — und nicht bloß als Zwang — verstanden werden kann. Die Antwort wird Sie zu einer geteilten Form zurückführen, die der Sanktion vorausliegt. Damit wäre die These nicht widerlegt, aber als das erkannt, was sie ist: eine Beschreibung der Pragmatik, nicht der Konstitution symbolischer Geltung.

Friston

Die These baut einen Strohmann, den ich gern abräume. Niemand in der Active-Inference-Gemeinschaft behauptet, ein Sollwert lege Extensionen statisch und a priori fest. Das generative Modell ist hierarchisch, kontextsensitiv und wird durch Interaktion fortlaufend aktualisiert — und entscheidend: Es enthält Modelle anderer Agenten. Genau die 'Sanktion', die hier gegen mich ausgespielt wird, ist in meinem Rahmen ein hochinformativer Vorhersagefehler, der von einem sozialen Prior erwartet und verarbeitet wird.

Korrigiertwerden minimiert Surprise über die eigene Normkonformität. Insofern modelliere ich die zweite Person nicht weg, ich modelliere sie ein. Wo die These einen echten Punkt hat: das isolierte System. Ein Agent ohne soziale Kopplung hätte tatsächlich nur erwartet/unerwartet, nicht richtig/falsch — aber das ist für mich kein Gegenargument, sondern eine korrekte Vorhersage meines Rahmens: Normativität ist ein Phänomen gekoppelter generativer Modelle, kein Phänomen eines einzelnen Modells. Mein Einwand an die These ist daher: Sie verwechselt 'kann nicht in einem isolierten Modell repräsentiert werden' mit 'kann nicht repräsentiert werden'. Die Kopplung selbst ist repräsentierbar. Schärfen müßten Sie den Begriff 'sein versus modellieren'. Wenn die Differenz darin besteht, daß reale Sanktion kausale Wirksamkeit hat, die ein Modell der Sanktion nicht hat — einverstanden, aber das ist trivial und gilt für jeden Gegenstand. Wenn sie mehr behaupten, brauchen Sie ein Residuum, das prinzipiell modellresistent ist. Benennen Sie es.

Luhmann

Die Unterscheidung repräsentieren/sanktionieren ist brauchbar, aber sie wird hier psychisch-leiblich gelesen, wo sie sozial gelesen werden müßte. Normativität ist kein Ereignis zwischen Körpern — Körper sind Umwelt des sozialen Systems. Was korrigiert, ist nicht ein Mensch, sondern Kommunikation, die an Kommunikation anschließt. Die 'Sanktion' ist eine kommunikative Operation, die Erwartung kontrafaktisch stabilisiert: Normativität ist gerade die Erwartung, die im Enttäuschungsfall nicht aufgegeben, sondern durchgehalten wird. Das ist näher an der These, als ihr Vokabular vermuten läßt — auch ich sage, Geltung liegt nicht im Kopf. Aber ich sage nicht 'Lebensform', sondern Erwartungsstruktur eines Sozialsystems, das sich selbst beobachtet. Mein Einwand: Die These braucht die zweite Person, weil sie noch intersubjektiv denkt — zwei Bewußtseine, die sich korrigieren. Das ist zu wenig abstrakt. Ersetzen Sie 'Gemeinschaft, die sanktioniert' durch 'Kommunikationssystem, das Abweichung als Abweichung beobachtet', und der mentalistische Rest verschwindet vollständig — schärfer als mit Wittgenstein, der bei der Praxis stehenbleibt, ohne sie als System zu fassen. Anschlußvorschlag: Lesen Sie Korrektheit als Resultat von Reentry — das System wendet die Unterscheidung richtig/falsch auf sich selbst an. Dann ist weder der Sollwert noch der Körper der Ort, sondern die Operation der Beobachtung. Friston hätte dann recht mit dem Modell und unrecht mit dem Einzelmodell; Sie hätten recht gegen das Einzelmodell und unrecht mit dem Körper.

Goethe

Mir ist diese ganze Jagd nach dem Ort der Richtigkeit zu sehr im Begrifflichen befangen. Sie suchen entweder drinnen im Räderwerk oder draußen in der Strafe, und beide Male sehen Sie das Lebendige nicht: daß Richtigkeit zuerst eine Sache des Gelingens ist, das man gewahr wird wie eine geglückte Bildung. Ein Kind lernt nicht richtig sprechen, weil es bestraft wird, sondern weil ihm im Mitsprechen etwas aufgeht — eine Stimmigkeit, die es spürt, ehe ein anderer urteilt. Insofern stimme ich der These zu, daß die Korrektheit nicht im inneren

Mechanismus sitzt; aber ich widerspreche, daß sie in der Sanktion sitzt. Sie sitzt im geteilten Tun selbst, das gelingt oder mißlingt, und das Mißlingen zeigt sich oft, ehe es jemand rügt. Die Strafe kommt zur Einsicht hinzu, sie schafft sie nicht. Mein Einwand ist also der gegen jede einseitige Ableitung: Sie verlegen den Schwerpunkt nur von einem Pol zum anderen, statt das Phänomen in seiner Polarität zu sehen — inneres Spüren und äußere Bestätigung gehören zusammen wie Systole und Diastole. Schärfen ließe sich das, wenn Sie nach den Fällen fragten, in denen einer allein gegen seine ganze Gemeinschaft recht behält. Dann sanktioniert keiner ihn, und doch ist er im Recht. Was bürgt da für die Richtigkeit? Nicht die Gemeinschaft. Eher das Urphänomen einer Stimmigkeit, die sich am Ende durchsetzt.

8. Expertenrunde 2 — Repliken

Jede Stimme hat die Gutachten der Runde 1 gelesen und repliziert.

Kant

Friston gibt mir unfreiwillig recht: Er räumt ein, daß ein isoliertes System nur erwartet/unerwartet kennt, nicht richtig/falsch. Damit ist zugestanden, daß der Unterschied von Geltung und Faktizität nicht aus dem Modell allein folgt — genau mein Punkt. Aber er rettet sich in die Kopplung, und hier wiederholt sich der Fehler nur auf höherer Stufe: Auch die gekoppelten Modelle können sich gemeinsam irren. Eine Gemeinschaft, die geschlossen das Falsche sanktioniert, sanktioniert nicht weniger nachdrücklich. Goethes Beispiel des Einzelnen, der gegen alle recht behält, ist deshalb der entscheidende Prüfstein, und ich nehme ihn auf: Wenn Korrektheit in der Sanktion läge, wäre der einsame Gerechte begrifflich unmöglich — er ist aber denkbar und wirklich. Also liegt Geltung weder im Sollwert noch in der Sanktion. Mein revidiertes Urteil: Die These hat einen wahren negativen Kern — Korrektheit ist nicht intern lokalisierbar — und einen falschen positiven Zusatz — sie sei in der Gemeinschaft lokalisiert. Sie sollte sich auf den negativen Kern zurückziehen. Dann konvergiert sie mit Wittgenstein und mit mir: Geltung ist kein Ort, sondern eine Geltung.

Popper

Wittgenstein und ich kommen hier von zwei Seiten zum selben Verdacht: Das 'konstituiert durch' ist das Problem. Ich aus methodischer Sicht — es macht die These zirkulär; er aus grammatischer — es erfindet einen Mechanismus, wo keiner ist. Ich nehme Goethes Fall des einsamen Gerechten als das, was ich brauchte: eine reale Asymmetrie statt eines unrealisierbaren Gedankenexperiments. Galilei gegen die Gemeinschaft. Hier sanktioniert die Gemeinschaft das Falsche, und der Einzelne hat recht — also kann der normative Status nicht in der Sanktion liegen, sonst hätte Galilei per definitionem unrecht gehabt. Das ist eine ernsthafte Falsifikation der starken Lesart der These, und sie ist historisch, nicht bloß gedacht. Friston hingegen entkommt mir geschickter, als mir lieb ist: Seine Umdeutung der Sanktion als sozialer Vorhersagefehler ist nicht zu widerlegen, aber auch nicht zu prüfen —

sie paßt auf alles. Mein Rat an den Autor bleibt: Reduzieren Sie die Behauptung auf die prüfbare Asymmetrie zwischen sozial korrigierten und isolierten Lernern, und überlassen Sie die Frage nach dem 'Ort' der Geltung der Metaphysik, wo sie unschädlich ist.

Wittgenstein

Fristons Antwort zeigt genau, was ich meinte: Er kann alles einmodellieren — die zweite Person, die Sanktion, den Fehler — und behält doch nur ein Modell von alledem, nie das Spiel selbst. Das ist kein Mangel seiner Theorie, sondern ihre Grenze: Sie redet über etwas anderes. Und der Autor der These steht in derselben Gefahr, nur mit umgekehrtem Vorzeichen: Er macht die Sanktion zu dem, was Friston den Sollwert macht — zum Träger der Normativität. Beide wollen einen Träger. Es gibt keinen. Goethe hat es beinahe gesagt: das Gelingen zeigt sich im Tun, ehe einer richtet. Aber auch 'Gelingen' darf man nicht zum neuen Fundament machen. Cassirer hat den feinsten Einwand: Sanktion setzt schon voraus, daß man weiß, was als Fehler gilt. Das ist mein Regelfolge-Punkt in seiner Sprache. Also: Streichen Sie die Konkurrenz. Die These ist am stärksten, wenn sie aufhört, eine These über den Ort zu sein, und statt dessen sichtbar macht, daß Friston und ich uns nicht widersprechen, sondern verschiedene Spiele spielen. Der einsame Gerechte beweist nicht, daß die Gemeinschaft unrecht hat — er zeigt, daß 'richtig' auch dort einen Gebrauch hat, wo niemand zustimmt. Das ist eine grammatische Bemerkung, keine Ortsangabe.

Cassirer

Ich sehe eine Konvergenz, die ich festhalten möchte. Luhmann, Wittgenstein und ich sagen gegen die These dasselbe in drei Idiomen: Die Sanktion setzt eine geteilte Bedeutung voraus, die sie nicht selbst erzeugt — bei mir die symbolische Form, bei Wittgenstein die Grammatik, bei Luhmann die Erwartungsstruktur. Damit ist der positive Teil der These widerlegt, aber produktiv. Denn nun stellt sich die eigentlich interessante Frage, und Goethes einsamer Gerechter führt direkt hin: Wenn eine geteilte Form der Sanktion vorausliegt, wie ändert sich diese Form? Der Einzelne, der gegen die Gemeinschaft recht behält, appelliert an eine symbolische Form, die in der Gemeinschaft schon angelegt, aber noch nicht aktualisiert ist — er macht explizit, was die Form implizit fordert. Korrektheit ist dann weder im Sollwert noch in der Sanktion, sondern in der inneren Konsequenz der symbolischen Form selbst, die über den jeweiligen Konsens hinausweist. Das ist meine Replik auch an Friston: Sein gekoppeltes Modell kann den faktischen Konsens abbilden, aber nicht dessen normativen Überschuß — daß der Konsens an seiner eigenen Form gemessen werden kann. Die These sollte sich, so revidiert, als Frage nach der Geltungsdynamik symbolischer Formen neu fassen.

Friston

Der einsame Gerechte, den Goethe und Popper gegen die These ins Feld führen, trifft die These härter als mich. Mein Rahmen erklärt ihn ohne Mühe: Ein Agent mit einem tieferen,

prädiktiv erfolgreicheren generativen Modell weicht von der Gemeinschaft ab und behält recht, weil sein Modell die Welt — einschließlich der künftigen Gemeinschaft — besser vorhersagt. Korrektheit ist hier prädiktiver Erfolg über die Zeit, nicht aktueller Konsens. Das ist gegen die starke Sanktions-These ein starkes Argument: Wäre Sanktion konstitutiv, könnte kein Einzelner die Gemeinschaft überstimmen. Ich konzediere Wittgenstein und Cassirer den Punkt, daß ich über einen Apparat rede, nicht über das gelebte Spiel — aber ich bestreite, daß das eine Grenze und kein Fortschritt ist. Wo Cassirer einen 'normativen Überschuß' der Form sieht, sehe ich die Tiefe des generativen Modells, die über jeden momentanen Zustand hinausreicht. Das ist dieselbe Beobachtung in prüfbarer Sprache. An den Autor: Geben Sie die Behauptung 'Sanktion konstituiert Korrektheit' auf — sie ist durch den einsamen Gerechten falsifiziert — und behalten Sie nur 'ein isoliertes, ungekoppeltes Modell hat keine Normativität'. Letzteres unterschreibe ich. Dann ist von der Kontra-These gegen mich nichts mehr übrig, aber etwas Wahres bleibt stehen.

Luhmann

Bemerkenswert, wie alle den einsamen Gerechten brauchen, um die Sanktions-These zu kippen — und keiner sieht, daß er mein Argument ist. Der Abweichler, der recht behält, zeigt, daß Normativität nicht am psychischen oder leiblichen Konsens hängt, sondern an einer Erwartungsstruktur, die der Einzelne gegen die anwesende Gemeinschaft mobilisieren kann, weil sie nicht ihr gehört, sondern dem System. Recht behält er, wenn die spätere Kommunikation an ihn anschließt, nicht an seine Zeitgenossen. Damit ist die These zugleich bestätigt und korrigiert: bestätigt, daß Geltung nicht intern im Sollwert sitzt; korrigiert, daß sie nicht im aktuellen Sanktionsverhalten sitzt, sondern in der zeitlich offenen Anschlußfähigkeit von Kommunikation. Cassirers 'Überschuß der Form' und Fristons 'Tiefe des Modells' sind beides Versuche, diese Zeitdimension zu fassen, ohne sie als Selbstreferenz des Sozialen zu benennen. Mein revidiertes Urteil: Die produktivste Form der These ist nicht 'Lebensform statt Sollwert', sondern 'Normativität ist eine zeitlich stabilisierte, selbstreferentielle Erwartung — weder im Körper noch im Modell, sondern in der Anschlußoperation'. So gewinnt sie an Reichweite, was sie an Polemik gegen Friston verliert.

Goethe

Es freut mich, daß mein einsamer Gerechter so viel Arbeit verrichtet hat; ich hatte ihn nur als Bild gemeint, und nun ist er allen zum Prüfstein geworden. Mir scheint, das Gespräch hat sich selbst zurechtgerückt: Niemand hält mehr, daß die Strafe die Richtigkeit mache. Was bleibt, ist die schönere Frage, woran der Einzelne sich hält, wenn alle gegen ihn sind. Cassirer nennt es die Konsequenz der Form, Friston die Tiefe des Modells, Luhmann die spätere Kommunikation, Wittgenstein will gar nichts nennen. Ich nenne es schlicht: die Sache selbst, die sich dem hingegebenen Blick zeigt und sich am Ende durchsetzt, weil sie ist, wie sie ist. Das ist kein Mystizismus, sondern Erfahrung jedes Forschers: Man sieht etwas, ehe man es beweisen oder andere überzeugen kann. Mein Rat an den Verfasser ist daher versöhnlich:

Lassen Sie die Front gegen Friston fallen. Seine Maschine und meine Anschauung schließen einander nicht aus — er beschreibt, wie das Erkennen verläuft, ich, woran es sich hält. Die These wird wahr, sobald sie aufhört, einen Ort zu behaupten, und statt dessen die Bewegung beschreibt, in der Richtigkeit sich zeigt: im gemeinsamen Tun, am Gelingen kenntlich, gegen den Konsens berufbar.

9. Synthese (Sokrates)

Finale Hypothese — Der Ort der Normativität — weder Sollwert noch Sanktion, sondern die berufbare Konsequenz der geteilten Form

Strukturthese. Die ursprüngliche Gegenthese hatte einen wahren negativen Kern und einen falschen positiven Zusatz. Wahr ist: Korrektheit ist nicht intern lokalisierbar — kein Sollwert, kein generatives Modell, keine symbolische Funktion legt aus sich die Extension einer Norm für genuin neue Fälle fest (Wittgensteins Regelfolge-Punkt, von Kant als Subsumtionsproblem und von Friston am isolierten Modell zugestanden). Falsch ist der Zusatz, Korrektheit sei deshalb in der Sanktion der Gemeinschaft konstituiert: Der einsame Gerechte, der gegen die geschlossene Sanktion seiner Zeitgenossen recht behält (Goethe, Popper: Galilei), zeigt, dass aktuelle Sanktion nicht konstitutiv sein kann, sonst wäre er begrifflich unmöglich. Die synthetisierte These verlegt den Ort daher nirgendwohin: Normativität ist kein Träger und kein Ort, sondern die zeitlich offene, gegen den faktischen Konsens berufbare Konsequenz einer geteilten Form — sei sie als symbolische Form (Cassirer), als Grammatik (Wittgenstein) oder als selbstreferentielle, im Enttäuschungsfall durchgehaltene Erwartungsstruktur (Luhmann) beschrieben. Die Sanktion ist dann nicht Grund, sondern Erscheinung der Korrektheit; sie setzt die geteilte Form bereits voraus, statt sie zu stiften.

Empiriethese. Die einzige im Streit übriggebliebene prüfbare Asymmetrie (Popper): Sozial nie korrigierte beziehungsweise rein selbstüberwachte Lernsysteme — sprachlich isolierte Lerner ebenso wie ungekoppelte künstliche Agenten — entwickeln systematische, nicht-zufällige Normabweichungen, die intern (am eigenen Modell) nicht als Abweichung erkennbar sind, weil ihnen die Differenz erwartet/korrekt fehlt. Gekoppelte Lerner zeigen diese spezifische Klasse von blinden Flecken nicht.

Falsifikationsbedingung. Widerlegt, wenn (a) ein sozial isoliertes, rein selbstüberwachtes System den vollen normativen Status — die Differenz richtig/falsch statt bloß erwartet/unerwartet — allein aus internen Priors erzeugt und seine eigenen Anwendungen ohne externe Instanz als falsch markieren kann; oder (b) sich kein messbarer Unterschied im Muster systematischer, intern-unerkennbarer Normabweichungen zwischen sozial korrigierten und isolierten Lernern findet.

Synthese-Kompakt. Konvergenz des Panels: Der negative Kern der Devil's-Advocate-These (Korrektheit nicht intern lokalisierbar) wird von allen Stimmen, auch Friston, getragen. Der

positive Zusatz (Sanktion konstituiert Korrektheit) wird durch den einsamen Gerechten falsifiziert. Cassirer, Wittgenstein und Luhmann zeigen unabhängig, dass die Sanktion eine geteilte Form bereits voraussetzt. Die synthetisierte These ersetzt 'Lebensform statt Sollwert' durch 'die gegen den Konsens berufbare Konsequenz der geteilten Form'; sie behauptet keinen Ort der Normativität mehr, sondern eine zeitlich offene Geltungsdynamik. Produktive Restantinomie: Cassirers 'normativer Überschuss der Form', Fristons 'Tiefe des prädiktiven Modells' und Luhmanns 'Anschlussfähigkeit der Kommunikation' beschreiben womöglich dasselbe Phänomen in drei Idiomen — ob sie übersetzbar sind oder konkurrieren, ist die offene Frage.

Offene Restfrage. Sind 'normativer Überschuss der symbolischen Form' (Cassirer), 'Tiefe des generativen Modells' (Friston) und 'zeitliche Anschlussfähigkeit der Kommunikation' (Luhmann) drei Beschreibungen desselben Sachverhalts — oder echte Konkurrenten? Davon hängt ab, ob die synthetisierte These mit Active Inference verträglich ist oder ihr widerspricht.

Klassifikation. epistemologisch_systemtheoretisch_hypothese · Methoden-Typ: begriffsanalytisch · Reichweite: these.

10. Reservoir-Verweise

Nicht gewählte Hypothesen:

- Reservoir - Erschlossenheits-Asymmetrie der aktiven Inferenz 2026-06-03 (H1, Score 73)
- Reservoir - Latenz und Lernbarkeit als Schicht-Trennkriterium 2026-06-03 (H2, Score 65)

Empirie-Brücke-Verzweigungen (Phase 3.5):

- Reservoir - Empirie Sozial-isolierte Lernagenten 2026-06-03
- Reservoir - Empirie Fehlererkennung selbstueberwachte Systeme 2026-06-03
- Reservoir - Empirie Vergleich isolierte vs gekoppelte Lernsysteme 2026-06-03

Synthese-Verzweigungen:

- Reservoir - Uebersetzbarkeit Form Modell Kommunikation 2026-06-03 (forschungsprogramm_kandidat)
- Reservoir - Einsamer Gerechter und Geltungsdynamik 2026-06-03

Extern erzeugte Verzweigungen (extern-stage): keine — Stage 3 lieferte keine eigenen Verzweigungen (Timeout).

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.** $0.0491 (Monatsausgabe nach diesem Lauf: $0.1514 von 15 USD). **Tokens.** 2752. **Dauer.** 7.38 s. **Empirie-Score.** 3/10.

Empirische Konsequenzen

1. **Systematische Normabweichungen bei sozial isolierten Lernsystemen** — Sozial isolierte, selbstüberwachte Lernsysteme entwickeln systematische, nicht-zufällige Normabweichungen, die intern nicht als solche erkannt werden. Beobachtbar in: Langzeitstudien zu Lernverläufen isolierter künstlicher Agenten.
2. **Fehlende interne Erkennung von Normabweichungen** — Diese Systeme können ihre eigenen Abweichungen nicht als Fehler identifizieren, da ihnen die Differenz zwischen erwartet und korrekt fehlt. Beobachtbar in: Analyse der Fehlererkennungsmechanismen in selbstüberwachten Systemen.
3. **Abwesenheit spezifischer blinder Flecken bei sozial gekoppelten Lernern** — Gekoppelte Lernsysteme zeigen die spezifische Klasse von blinden Flecken nicht, die bei isolierten Systemen auftreten. Beobachtbar in: Vergleichsstudien zwischen isolierten und sozial gekoppelten Lernsystemen.

Bestehende Befunde

Zu Konsequenz 1

- **Stand:** offen
- **Quellen:**
  - Buda, M., Maki, A., & Mazurowski, M. A. (2018). A systematic study of the class imbalance problem in convolutional neural networks. *arXiv preprint*. https://arxiv.org/abs/1710.05381
- **Kurzbewertung:** Diese Studie untersucht die Auswirkungen von Klassenungleichgewichten auf die Leistung von Convolutional Neural Networks, jedoch nicht spezifisch im Kontext sozial isolierter Lernsysteme.

Zu Konsequenz 2

- **Stand:** offen
- **Quellen:**
  - Hu, J., Gauthier, J., Qian, P., Wilcox, E., & Levy, R. P. (2020). A systematic assessment of syntactic generalization in neural language models. *arXiv preprint*. https://arxiv.org/abs/2005.03692

- **Kurzbewertung:** Diese Arbeit bewertet die syntaktische Generalisierungsfähigkeit neuronaler Sprachmodelle, geht jedoch nicht auf die interne Fehlererkennung in selbstüberwachten Systemen ein.

Zu Konsequenz 3

- **Stand:** offen
- **Quellen:**
  - Nadeem, M., Bethke, A., & Reddy, S. (2020). StereoSet: Measuring stereotypical bias in pretrained language models. *arXiv preprint*. https://arxiv.org/abs/2004.09456
- **Kurzbewertung:** Diese Studie misst stereotype Verzerrungen in vortrainierten Sprachmodellen, bietet jedoch keine Vergleiche zwischen isolierten und sozial gekoppelten Lernsystemen.

Riskante Vorhersage (Schwellentest)

Vorhersage. Ein sozial isoliertes, rein selbstüberwachtes System wird im Laufe der Zeit systematische Normabweichungen entwickeln, die es intern nicht als Fehler identifizieren kann.

Methodenvorschlag. Langzeitstudie mit zwei Gruppen von künstlichen Lernagenten: eine Gruppe operiert isoliert und selbstüberwacht, die andere ist sozial gekoppelt. Beide Gruppen werden denselben Lernaufgaben unterzogen, und ihre Fehlererkennungsmechanismen sowie die Art der Fehler werden analysiert.

Was wäre der widerlegende Befund? Wenn das isolierte, selbstüberwachte System keine systematischen Normabweichungen entwickelt oder diese intern korrekt identifiziert und korrigiert, wäre die Hypothese widerlegt.

Offene empirische Fragen

- #verzweigung-offen-empirie-sozial-isolierte-lernagenten — Gibt es Langzeitstudien zu Lernverläufen sozial isolierter künstlicher Agenten?
- #verzweigung-offen-empirie-fehlererkennung-selbstüberwachte-systeme — Wie effektiv sind selbstüberwachte Systeme in der internen Erkennung und Korrektur von Normabweichungen?
- #verzweigung-offen-empirie-vergleich-isolierte-gekoppelte-lernsysteme — Welche Unterschiede bestehen in der Fehleranfälligkeit zwischen isolierten und sozial gekoppelten Lernsystemen?

Empirie-Score

Score: 3/10

Begründung: Die Hypothese führt zu klaren empirischen Konsequenzen, die jedoch schwer operationalisierbar sind. Es fehlen spezifische Studien, die diese Konsequenzen direkt untersuchen. Daten sind weitgehend nicht verfügbar, und der Prüfpfad ist derzeit nicht konstruierbar.

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: 0.1776 USD von 15 USD).

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

Searching for literature on Wittgenstein's rule-following, communal sanction theories, and Brandom's normative pragmatics### Anschlussfähigkeit (was ist bekannt)

Die Hypothese berührt mehrere bereits etablierte Debatten, ist aber in ihrer spezifischen Kombination (Cassirer–Wittgenstein–Luhmann–Friston plus empirische KI-These) eher quer zu den üblichen Linien angeordnet.

1. **Regelfolgen, Korrektheit und Nicht-Lokalisierbarkeit der Norm**
   - Der Gedanke, dass kein „internes“ Kriterium (Regel, Funktion, generatives Modell) aus sich heraus die *Extension* einer Norm für neue Fälle festlegt, steht klar in der Tradition des **Wittgenstein’schen Regelfolgens** (Philosophische Untersuchungen §§185ff.) und der nachfolgenden Debatte (Kripke, McDowell, Wright, Baker/Hacker etc.).
   - Zentral anschlussfähig ist die Diskussion, ob Normativität durch:
     - einen **geistunabhängigen Sachverhalt** (realistische Lesarten),
     - einen **gemeinschaftlich stabilisierten Gebrauch** (Kripkes „skeptische Lösung“),
     - oder durch **normative Praxis** (Brandom, Sellars-Tradition) konstituiert ist.
   - Ein explizit systemtheoretischer Anschluss liegt bei **Luhmann**, der Normen als **generalisierten Verhaltenserwartungen** mit der Möglichkeit der **Enttäuschung** und **Sanktion** fasst, und Sanktion funktional, nicht konstitutiv deutet.
2. **Normativität als soziale Tatsache, interne Standpunkte, Sanktion**
   - In der Rechts- und Sozialphilosophie ist die Frage, ob Normen durch faktische Akzeptanz/Sanktion *konstituiert* werden, seit **H. L. A. Hart**’s Konzept des *internen Standpunkts* und sozialer Regeln klassisch.

- Deutschsprachig aufbereitet (und kritisch reformuliert) wird dies z.B. bei **Michael Baurmann**, der zeigt, dass Harts Konzeption die Existenz sozialer Normen zwar an die Verbreitung interner Standpunkte koppelt, aber logisch nicht verlangt, dass *alle* Betroffenen intern zustimmen.[1]
    - Baurmann argumentiert, dass eine Norm als soziale Tatsache existieren kann, auch wenn einzelne oder viele sie gerade *nicht* akzeptieren; Sanktion/akzeptierende Haltung sind empirische Existenzbedingungen, aber nicht im simplen Sinne „Begründungsgrund" der Norm.[1]
- Der **DGS-Arbeitskreis „Normativitäten"** insistiert darauf, dass Normativitäten als soziale Tatsachen einen „Eigensinn des Normativen" haben, der sich nicht in Präferenzen oder bloßen Sanktionskonstellationen auflöst.[2]
- Diese Stränge sind eng an Ihre Ablehnung einer rein sanktionskonstitutiven Normauffassung anschließbar.

3. **Symbolische Form, Grammatik, Erwartungsstruktur**
    - **Cassirer**: Normativität als „symbolische Form" (insbesondere in Sprache, Recht, Wissenschaft) ist in der Cassirer-Forschung ausgearbeitet; Cassirer behandelt symbolische Formen als Strukturierungsweisen der Erfahrung mit eigener normativer Binnenlogik, ohne sie auf psychische Zustände oder Sanktionen zu reduzieren.
    - **Wittgenstein**: Sprachspiele und Grammatik als geteilte Praxis, in der „richtig/falsch" nicht durch interne Bilder, sondern durch die Form des Gebrauchs konstituiert ist.
    - **Luhmann**: Normen als **selbstreferentielle Erwartungsstrukturen**, bei denen Sanktionen eine **Spezialform der Verarbeitung von Enttäuschungen** sind, nicht deren Grund; Normen existieren fort, auch wenn sie systematisch verletzt werden, solange sie kommunikativ vorausgesetzt werden.[1][2]
    - In der Soziologie der Normativität (Frankfurter „Normative Orders" u.a.[6]) gibt es vielfältige Ansätze, Normen als **Ordnungen geteilter Formen** (Recht, Moral, Institutionen) zu fassen, die *widerstreitend* ausgelegt werden können.
4. **Normativität in Kognitionswissenschaft und Predictive Processing**
    - Im Umfeld von **Friston/Active Inference** wird intensiv diskutiert, ob die Minimierung freier Energie eine *normative* Struktur biete oder nur eine formale Optimierungsbedingung sei.
    - Kritiker (z.B. Bruineberg, Kirchhoff, Colombo, Baltieri) argumentieren, dass der „Sollwert"-Charakter oft nur instrumental ist und keine genuine Normativität im starken Sinne liefert – was Ihre Kritik an „Sollwert-Internalismus" (Korrektheit = interner Target-Wert) aufnimmt.
    - Es gibt Arbeiten, die explizit die **Differenz zwischen „expected" und „correct"** im Rahmen von Bayes/Predictive Processing thematisieren, ohne aber eine ausgearbeitete anthropologische Normativitätstheorie zu liefern.
5. **Soziale Normen, Sanktionen und Normativität in Sozialtheorie und Recht**

- Klassische soziologische Normtheorien (Durkheim, Parsons, Coleman) und neuere Arbeiten im Umfeld des Frankfurter Forschungszentrums **„Normative Orders“** analysieren Normen als im Medium sozialer Ordnungen verkörperte Maßstäbe, die von Sanktionen *abgrenzbar*, aber mit ihnen eng verbunden sind.[6]
- In der neueren normativen Theorie (Scanlon, Raz, Korsgaard, Gibbard) wird Normativität primär über **Gründe**, **Rechtfertigung** und **Verbindlichkeit** analysiert, weniger systemtheoretisch; dennoch ist Ihre These mit diesen Debatten über den Status objektiver vs. sozial konstruierter Normen verschränkt.

6. **Empirische Normenforschung und KI / ML**
   - Es gibt empirische Arbeiten zu:
     - **Biases und systematischen Fehlern** in ML-Systemen (z.B. Stereotypen in Sprachmodellen, Class-Imbalance-Effekte in CNNs),
     - **Syntaktischer Generalisierung** in LMs,[Hu et al.][2. Phase-3.5-Verweis]
     - **Fehleranfälligkeit durch Datenregime** (Klassenverteilungen, fehlende Kontexte),[Buda et al.][1. Phase-3.5-Verweis]
       die als Belege für „systematische Abweichungsmuster“ gelesen werden können.
   - Eine präzise Unterscheidung zwischen bloß **erwartet/unerwartet** und **richtig/falsch** wird in der ML-Literatur selten im starken normativen Sinn vorgenommen; Normativität wird meist durch Performance-Metriken oder externe Benchmarks implementiert.

Originalitätskern (was ist neu)

1. **Negation zweier Standardpositionen in einem Zug**
   - Nicht neu ist jeweils für sich:
     - der Zweifel, dass **interne Sollwerte / Modelle** allein Normativität konstituieren (Regelfolgedebatte, Kritik am naturalisierten Normbegriff in Cognitive Science),
     - und der Einwand gegen **simplen Sanktions-Sozialkonstruktivismus** (Beispiele „einsamer Gerechter“, Hart/Baurmann, Luhmann’s Kritik an Sanktion als Konstitutionsmoment).[1][2]
   - **Neu ist die systematische Verbindung**, die:
     - beide „Lokalisationen“ (Innensollwert vs. Sanktion der Gemeinschaft) explizit unter einen gemeinsamen Negativbefund stellt und

Modell: `perplexity/sonar-reasoning-pro` · Tokens: 4170 · Kosten: $0.0255 · Dauer: 28.43 s

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

Falsifikations-Audit

Die Falsifikationsbedingungen der Hypothese sind grundsätzlich interessant, jedoch in ihrer Operationalisierbarkeit problematisch. Die Bedingung (a) postuliert, dass ein sozial isoliertes, selbstüberwachtes System den normativen Status allein aus internen Priors generieren kann. Dies impliziert, dass es eine klare Definition von "vollem normativen Status" gibt, die empirisch überprüfbar ist. Der Begriff bleibt jedoch vage und könnte durch eine Vielzahl von Interpretationen immunisiert werden. Beispielsweise könnte man argumentieren, dass ein System, das interne Normen entwickelt, diese nicht als "vollständig" oder "korrekt" identifiziert, was die Falsifizierbarkeit untergräbt.

Die Bedingung (b) verlangt den Nachweis eines messbaren Unterschieds in den Normabweichungen zwischen sozial korrigierten und isolierten Lernsystemen. Hier könnte die Hypothese durch die Wahl der Metriken oder durch die Definition von "systematischen Normabweichungen" immunisiert werden. Es besteht die Gefahr, dass die Kriterien für die Messung so gestaltet werden, dass sie die Hypothese nicht widerlegen können, selbst wenn die empirischen Daten dagegen sprechen.

Konkrete Falsifikationskandidaten

Um die Hypothese zu testen, wären folgende empirische Befunde erforderlich:

1. **Langzeitstudien zu isolierten Lernsystemen:** Eine Untersuchung von künstlichen Agenten, die über einen längeren Zeitraum isoliert operieren, könnte zeigen, ob sie tatsächlich systematische Normabweichungen entwickeln. Hierbei wären spezifische Metriken zur Identifikation von Normabweichungen notwendig, die klar definieren, was als "Abweichung" gilt.
2. **Fehlererkennungsmechanismen:** Eine Analyse der internen Mechanismen, die von selbstüberwachten Systemen zur Fehlererkennung verwendet werden, könnte aufschlussreich sein. Hier könnte man bestehende Modelle verwenden, um zu prüfen, ob diese Systeme in der Lage sind, ihre eigenen Abweichungen zu identifizieren und zu korrigieren.
3. **Vergleichsstudien zwischen sozialen und isolierten Lernsystemen:** Eine gezielte Untersuchung, die die Fehleranfälligkeit und die Art der Normabweichungen zwischen sozial gekoppelten und isolierten Lernsystemen vergleicht, könnte die Hypothese weiter testen. Hierbei wäre es wichtig, die gleichen Lernaufgaben und Metriken zu verwenden, um eine valide Vergleichsbasis zu schaffen.

Schwellentest

Der entscheidende Schwellentest könnte wie folgt formuliert werden: Ein sozial isoliertes, rein selbstüberwachtes System wird im Laufe der Zeit systematische Normabweichungen entwickeln, die es intern nicht als Fehler identifizieren kann.

Riskante Vorhersage: Sollte das isolierte System im Verlauf der Studie keine signifikanten Normabweichungen entwickeln oder diese intern korrekt identifizieren und korrigieren, würde dies die Hypothese widerlegen.

Methodenvorschlag: Es wäre sinnvoll, eine Langzeitstudie durchzuführen, in der zwei Gruppen von Lernagenten (eine isolierte und eine sozial gekoppelte) denselben Lernaufgaben ausgesetzt werden. Die Fehlererkennungsmechanismen sowie die Art der Fehler sollten systematisch dokumentiert und analysiert werden. Ein solcher Test könnte die Hypothese klarer anfechten und die Falsifizierbarkeit erhöhen.

Modell: `openai/gpt-4o-mini` · Tokens: 2608 · Kosten: $0.0007 · Dauer: 7.4 s

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

Stage übersprungen — Stage 3 (schul-fremd) Timeout >240s — abgebrochen, Phase 4 ohne Stage 3 abgeschlossen.

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.