Inhaltsverzeichnis

Hypothesentag-Gutachten

Die Gewinnerthese

Bindung unter durchschauter Versuchung

Initiale Fassung

Hypothese 1 (warm_pick: empirie-provenienz-pupillendilatation) — Pupille als privilegierter Präzisionsmarker

Kernsatz. Die Pupillendilatation ist im Provenienz-Paradigma kein beliebiger der drei online-Marker, sondern der privilegierte: Sie misst nicht generische Erregung, sondern direkt die Präzisionsgewichtung (LC-NE-vermittelter Gain), mit der ein Provenienz-Prior das prädiktive Modell des Betrachters umstrukturiert. Wenn die Synthese vom 08.06. recht hat, dass Provenienz als Präzisions-Prior wirkt, dann muss die Pupille — als bekannter Index des Präzisions-Gains — den Effekt *monoton mit der Reizmehrdeutigkeit* tragen, dort also am stärksten, wo das generative Modell am wenigsten festgelegt ist.

Begründung. Die offene Verzweigung `empirie-provenienz-pupillendilatation` aus [[06 Hypothesentag/2026-06-08]] fragte, ob die Pupille den Provenienzeffekt vor der verbalen Bewertung trägt. Die Synthese fasste Provenienz als Modulation der *Präzision* der Priors über die generative Ursache des Werks ([[Cassirer-Friston Doppelthese - Ermöglichung und strukturelle Homologie]]). Die Neurophysiologie liefert die Brücke: Pupillenweite ist ein etablierter Proxy des Locus-coeruleus-Noradrenalin-Systems, das im Predictive-Processing-Rahmen genau die Präzisionsgewichtung (Gain auf Vorhersagefehler-Einheiten) reguliert. Daraus folgt eine schärfere, riskantere Vorhersage als „die Pupille trennt“: Die Trennung der Pupillendynamik unter „Original” vs. „Fälschung” muss mit der strukturellen Mehrdeutigkeit des Reizes skalieren, weil Präzisionsgewichtung dort am meisten leistet, wo der Input unterbestimmt ist.

Falsifikationsbedingung. Widerlegt, wenn (a) die Pupillentrennung bei identischem Reiz unter

beiden Provenienz-Etiketten *flach* bleibt über das Mehrdeutigkeitskontinuum (kein monotoner Anstieg), oder (b) die ästhetisch irrelevante Kontroll-Provenienz (Herkunft eines Alltagsgegenstands) dieselbe mehrdeutigkeitsabhängige Pupillentrennung erzeugt — dann misst die Pupille generische Salienz, nicht symbolisch-funktionale Präzision. Marker, Zeitfenster und das Mehrdeutigkeitsmaß sind vor der Auswertung erschöpfend zu fixieren (Immunisierungssperre).

Quelle. warm_pick — verfolgt den Strang aus [[06 Hypothesentag/2026-06-08]] (Provenienz als Präzisions-Prior), vertiefend auf den schmalsten der drei online-Marker.

Reformulierte Fassung (nach Kritischem Professor)

Hypothese 1 (warm_pick: empirie-provenienz-pupillendilatation) — Pupille als privilegierter Präzisionsmarker

Kernsatz. Die Pupillendilatation ist im Provenienz-Paradigma kein beliebiger der drei online-Marker, sondern der privilegierte: Sie misst nicht generische Erregung, sondern direkt die Präzisionsgewichtung (LC-NE-vermittelter Gain), mit der ein Provenienz-Prior das prädiktive Modell des Betrachters umstrukturiert. Wenn die Synthese vom 08.06. recht hat, dass Provenienz als Präzisions-Prior wirkt, dann muss die Pupille — als bekannter Index des Präzisions-Gains — den Effekt *monoton mit der Reizmehrdeutigkeit* tragen, dort also am stärksten, wo das generative Modell am wenigsten festgelegt ist.

Begründung. Die offene Verzweigung `empirie-provenienz-pupillendilatation` aus [[06 Hypothesentag/2026-06-08]] fragte, ob die Pupille den Provenienzeffekt vor der verbalen Bewertung trägt. Die Synthese fasste Provenienz als Modulation der *Präzision* der Priors über die generative Ursache des Werks ([[Cassirer-Friston Doppelthese - Ermöglichung und strukturelle Homologie]]). Die Neurophysiologie liefert die Brücke: Pupillenweite ist ein etablierter Proxy des Locus-coeruleus-Noradrenalin-Systems, das im Predictive-Processing-Rahmen genau die Präzisionsgewichtung (Gain auf Vorhersagefehler-Einheiten) reguliert. Daraus folgt eine schärfere, riskantere Vorhersage als „die Pupille trennt“: Die Trennung der Pupillendynamik unter „Original” vs. „Fälschung” muss mit der strukturellen Mehrdeutigkeit des Reizes skalieren, weil Präzisionsgewichtung dort am meisten leistet, wo der Input unterbestimmt ist.

Falsifikationsbedingung. Widerlegt, wenn (a) die Pupillentrennung bei identischem Reiz unter beiden Provenienz-Etiketten *flach* bleibt über das Mehrdeutigkeitskontinuum (kein monotoner Anstieg), oder (b) die ästhetisch irrelevante Kontroll-Provenienz (Herkunft eines Alltagsgegenstands) dieselbe mehrdeutigkeitsabhängige Pupillentrennung erzeugt — dann misst die Pupille generische Salienz, nicht symbolisch-funktionale Präzision. Marker, Zeitfenster und das Mehrdeutigkeitsmaß sind vor der Auswertung erschöpfend zu fixieren (Immunisierungssperre).

Quelle. warm_pick — verfolgt den Strang aus [[06 Hypothesentag/2026-06-08]] (Provenienz als Präzisions-Prior), vertiefend auf den schmalsten der drei online-Marker.

Expertenrunden und Synthese

7. Bewertung mit dreifacher Klassifikation

Gewichte: Originalität 1.6 · Tiefe 1.2 · Antinomie-Test 1.2 · übrige sechs 1.0. Normfaktor 0.882353 (×, auf ganze Zahl gerundet).

H1 — Pupille als privilegierter Präzisionsmarker

Klassifikation: Themenfeld `epistemologisch_systemtheoretisch_hypothese` · Methoden-Typ `empirisch_pruefbar` · Reichweiten-Klasse `sondierung` (schmale Einzelaussage, vertieft den Provenienz-Strang).

Kriterium	Score	Begründung
Originalität	6	Verengung des bereits abgesegneten Provenienz-Strangs; Pupille-Präzisions-Link ist in der Pupillometrie angelegt.
Falsifizierbarkeit	9	Präregistriertes Design, Interaktionstest, Kontrollbedingung, Last-Herauspartialisierung — scharf und riskant.
Begriffliche Klarheit	8	„Privilegiert" erschöpfend als stärkste Mehrdeutigkeits-Skalierung operationalisiert.
Tiefe	5	Marker-Frage, kein Griff ins Grundlegende; bewusst schmal gehalten.
Forschungsrelevanz	8	Neuroästhetik und Predictive-Processing-Ästhetik sind aktive Felder.
Interdisziplinär	7	Neurophysiologie, Predictive Processing, Kunstphilosophie.
Vault-Anschluss	9	Direkter warm_pick, vertieft 06-08 und die Cassirer-Friston-Doppelthese.
Antinomie-Test	6	Gegenthese „Pupille misst generische Salienz" ist plausibel, aber durch Kontrollbedingung adressiert.
Publikationsmöglichkeit	8	Sauberes empirisches Paper-Design, klar publizierbar.

Gewichtete Summe → 63/90.

H2 — Der Narr als kostspieliger Erstabweichler

Klassifikation: Themenfeld `ethisch_praktische_hypothese` (sozialmechanisch) · Methoden-Typ `empirisch_pruefbar` (mit begriffsanalytischem Kern) · Reichweiten-Klasse `these`.

Kriterium	Score	Begründung
Originalität	7	Soziologische Umdeutung des Narr-Erlöser-Topos; der *vergleichende* Anspruch (stark vs. schwach fokale Felder) ist frisch.
Falsifizierbarkeit	7	Gini-Konzentrationsmaß + Kausalitätstest, aber Felddaten zu Sanktionslasten schwer erhebbar.
Begriffliche Klarheit	7	Schelling-Punkt und „gemeinsames Wissen" klar, „Erstabweichler" sauber definiert.

Kriterium	Score	Begründung
Tiefe	6	Greift den Henne-Ei-Mechanismus der Konventionsänderung, bleibt aber sozialtheoretisch mittlerer Reichweite.
Forschungsrelevanz	7	Normwandel- und Konventionsforschung (Bicchieri, Cultural Evolution) sind aktiv.
Interdisziplinär	8	Soziologie, Spieltheorie, Psychologie, Kulturevolution, Wissenschaftsgeschichte.
Vault-Anschluss	7	Virgin Node Narr/Erlöser + Reservoir Kanon-als-Schelling-Punkt.
Antinomie-Test	7	Granovetters Schwellen-Kaskaden-Modell ist eine starke, ernstzunehmende Gegenthese.
Publikationsmöglichkeit	7	Publizierbar, sofern Operationalisierung an einem konkreten Korpus gelingt.

Gewichtete Summe → 62/90.

H3 — Das unbedingte Sollen als Commitment-Strategie

Klassifikation: Themenfeld `ethisch_praktische_hypothese` · Methoden-Typ `empirisch_pruefbar` (mit normativ-begriffsanalytischem Kern; Genese-These, nicht Geltungs-These) · Reichweiten-Klasse `these`.

Kriterium	Score	Begründung
Originalität	8	Nicht „Moral als Commitment" (bekannt seit Frank), sondern speziell die *phänomenale Unbedingtheit* als das selektierte Signal — diese Verengung ist neu.
Falsifizierbarkeit	7	Vorab fixierter nicht-strategischer Kontrollfall + riskante Korrelationsvorhersage; ESS-Anspruch nur bedingt erhoben.
Begriffliche Klarheit	7	Genese/Geltung sauber getrennt, Kategorienfehler explizit vermieden.
Tiefe	8	Trifft die Fundierung der Normativität und den Status des Kantischen Apriori.
Forschungsrelevanz	8	Naturalisierung der praktischen Vernunft ist eine der lebendigsten metaethischen Debatten.
Interdisziplinär	9	Ethik, Metaethik, Spieltheorie, Evolutionspsychologie, Kant-Exegese.

Kriterium	Score	Begründung
Vault-Anschluss	7	Dispositionaler Rest der Moral, Evans-Kant-Formalisierung, Inversion der Moral.
Antinomie-Test	8	Die Antinomie Genese Geltung ist stark und produktiv — eine echte Spannung, nicht auflösbar.
Publikationsmöglichkeit	8	Als metaethisch-naturalistischer Beitrag mit empirischer Schiene gut platzierbar.

Gewichtete Summe → 69/90.

Auswahl

Hypothese	Summe
H3 — Das unbedingte Sollen als Commitment-Strategie	**69**
H1 — Pupille als privilegierter Präzisionsmarker	63
H2 — Der Narr als kostspieliger Erstabweichler	62

Gewählt: H3. Höchste gewichtete Summe, getragen von Originalität (8), Tiefe (8) und Antinomie-Test (8) — genau die hochgewichteten Kriterien. Sie bringt die Diversitäts-Pflicht des Tages (cold-Rotation aus Psychologie_Ethik/Spieltheorie) in die Gewinnerposition und durchbricht damit aktiv die Cassirer-Makro-Drift (80%).

Reichweiten-Klasse der Gewinnerin: `these`. Begründung: eigenständig prüfbare Behauptung mit weiter, aber fokussierter Reichweite (ein identifizierter Mechanismus: phänomenale Unbedingtheit = Commitment-Signal). Bewusst *nicht* zum Forschungsprogramm inflationiert — Monats-Review: forschungsprogramm 71% steril. Kein `forschungsprogramm_kandidat`.

Panel (ethisch_praktische_hypothese): Kant · Popper · Wittgenstein · Plessner · Aristoteles · Nietzsche · Hannah Arendt.

Reservoir-Profile (nicht gewählte Hypothesen)

H1 → Reservoir „Pupille Präzisionsmarker" · reichweiten_klasse: sondierung · forschungsprogramm_kandidat: false · Empfehlung: als empirische Verzweigung des Provenienz-Strangs reaktivieren, sobald ein Pupillometrie-Datensatz verfügbar ist. Offen: Interaktion Provenienz × Mehrdeutigkeit nach Last-Kontrolle.

H2 → Reservoir „Narr als Erstabweichler" · reichweiten_klasse: these · forschungsprogramm_kandidat: false · Empfehlung: aufgreifen mit einem konkreten Normwandel-Korpus (Sprachwandel oder Paradigmenwechsel), um Sanktions-Konzentration (Gini) gegen Granovetter-Kaskaden zu testen. Anschlussthese zur Schelling-Punkt-Linie im Vault.

8. Expertenrunde 1

Expertenrunde 1 — unabhängige Gutachten

Panel-Konfiguration. Themenfeld `ethisch_praktische_hypothese` (die These behauptet etwas über den Status moralischer Verpflichtung). Panel `ethisch_praktische_hypothese` gewählt.

Panel-Mitglieder. Kant, Popper, Wittgenstein, Plessner, Aristoteles, Nietzsche, Hannah Arendt.

Gutachten 1 — Kant

Die These trifft mein Herzstück, und gerade deshalb muss ich scharf trennen. Sie behauptet, der Charakter der Unbedingtheit sei die *Innenseite* einer Commitment-Strategie. Hier liegt ein Kategorienfehler in Lauerstellung, den die Reformulierung zwar benennt, aber nicht entschärft. Daß ein Pflichtgefühl evolutionär nützlich ist, sagt über seinen Geltungsanspruch nichts — das ist die genetische Täuschung. Mein eigentlicher Einwand ist tiefer: Wenn das „nicht anders Können" als selektierter Mechanismus beschrieben wird, dann ist es gerade *nicht* mehr das moralische Sollen. Denn das Sollen setzt Freiheit voraus — ich soll, *also kann ich auch nicht*. Ein Akteur, der die Defektionsoption verloren hat, handelt nicht aus Pflicht, sondern aus Naturnotwendigkeit; er ist heteronom bestimmt, ein Spielstein der Auszahlungsmatrix. Was die These als Stärke des Signals verkauft — die verlorene Wahlfreiheit —, ist moralphilosophisch die Abschaffung der Moral. Ich verlange daher die Präzisierung: Behauptet die These, das *Phänomen* der Unbedingtheit sei strategisch erzeugt (das ist prüfbar und vielleicht wahr), oder behauptet sie, die *Verbindlichkeit* selbst sei nichts als dieser Mechanismus (das ist falsch und überschreitet ihre Mittel)? Im ersten Fall beschreibt sie die Anthropologie des moralischen Gefühls; im zweiten verwechselt sie die Spur mit dem Gesetz. Der kategorische Rest bleibt: dasjenige, das *urteilt*, diese Bindung solle gelten, geht in keiner Strategie auf.

Gutachten 2 — Popper

Methodisch ist diese These besser gebaut als die meisten, die ich sehe — und genau darum prüfe ich sie streng. Das Lob zuerst: Der vorab fixierte, nicht-strategische Kontrollfall ist der richtige Zug. Ohne ihn wäre die These ein Konfirmationszirkel, denn jede Pflicht ließe sich nachträglich als „eigentlich strategisch" umdeuten. Die Reformulierung sperrt diese Umdeutung — das verlange ich, und das ist geliefert. Aber drei Lücken bleiben. Erstens: Was zählt operational als „rein nicht-strategisch"? Die Liste (keine Beobachtung, keine Reziprozitätserwartung) muss *erschöpfend* und vor der Erhebung deklariert sein, sonst schiebt man bei negativem Befund eine vergessene strategische Dimension nach. Zweitens: Die ESS-Behauptung ist modellabhängig. Frank und Axelrod liefern Existenzbeweise unter spezifischen Auszahlungsstrukturen — die These darf daraus keine Allaussage machen. Sie tut es klugerweise nicht, aber sie muss die Bedingungen *nennen*, unter denen sie falsch wäre. Drittens, mein konkreter Falsifikationskandidat: Wenn die erlebte Unbedingtheit bei Personen mit reduzierter strategischer Kognition (man denke an experimentelle Manipulation der Reziprozitätswahrnehmung) *unverändert* bleibt, ist sie kein Commitment-Device. Diesen Test würde ich ins Design aufnehmen. Solange er aussteht, bleibt die These eine gut gebaute Vermutung — nicht widerlegt, aber auch nicht gestützt.

Gutachten 3 — Wittgenstein

Welches Sprachspiel spielen wir, wenn wir sagen, jemand „könne nicht anders"? Die These behandelt diesen Ausdruck, als bezeichne er einen inneren Zustand — ein Gefühl der Unbedingtheit, das dann gemessen und mit einem Mechanismus identifiziert wird. Aber „ich kann nicht anders" ist im moralischen Sprachspiel selten ein Bericht über ein Erlebnis. Es ist ein Zug: Luther vor dem Reichstag berichtet nicht über seine Pupillen, er nimmt Stellung. Die Grammatik von „müssen" ist hier nicht die der Kausalität („der Stein muss fallen"), sondern die der Verpflichtung — und die beiden zu verwechseln ist die alte Krankheit. Die These droht, das normative „Ich muss" in ein psychologisches „Ich bin gezwungen" zu übersetzen und dann zu staunen, daß sie einen Mechanismus findet. Sie hat ihn hineingelegt. Mein Vorschlag ist nicht, die These zu verwerfen, sondern ihre Behauptung umzuformulieren: Nicht „die Unbedingtheit *ist* ein Commitment-Device", sondern „der Gebrauch unbedingter Verpflichtungsworte *funktioniert wie* eine Selbstbindung in bestimmten Praktiken — im Versprechen, im Schwur, im Bekenntnis." Das ist prüfbar am Verhalten und vermeidet die innere Bühne. Was das Versprechen bindet, zeigt sich darin, daß man es hält, nicht in einem verborgenen Gain-Parameter. Messt den Gebrauch, nicht das Gefühl.

Gutachten 4 — Plessner

Ich plädiere, wie so oft, für die schwächere Form. Die These hat recht in dem, was sie sieht: Verpflichtung hat eine Tiefe, die sich nicht in Kalkül auflöst, und sie wirkt sozial bindend. Aber sie verfehlt das Phänomen, wenn sie das Erleben der Unbedingtheit zur *Operation* eines Mechanismus macht. Verpflichtetsein ist eine Stellung, keine Operation. Der Mensch steht exzentrisch zu seinen eigenen Bindungen: Er kann sich unbedingt verpflichtet *fühlen* und im selben Atemzug *wissen*, daß diese Bindung kontingent, erworben, vielleicht strategisch grundiert ist — und sie hält dennoch. Genau diese Doppelung ist der anthropologische Ort, den die These übergeht. Ein bloßes Commitment-Device kennt diese Spannung nicht; es bindet oder bindet nicht. Der Mensch aber kann seine Bindung durchschauen und sie trotzdem vollziehen — das ist keine Naturnotwendigkeit, sondern eine Leistung der Distanz zu sich selbst. Mein Schärfungsvorschlag: Die These

solle „die Unbedingtheit *ist* die Innenseite einer Strategie" abschwächen zu „strategische Verwundbarkeit ist eine *notwendige, nicht hinreichende* Bedingung dafür, daß sich unbedingte Verpflichtung ausbildet." Dann bleibt Raum für den Fall, der die reine Mechanik sprengt: daß einer sich gebunden weiß, wo niemand zusieht und nichts zu gewinnen ist — nicht weil er die Lage falsch einschätzt, sondern weil er zu sich selbst Stellung genommen hat.

Gutachten 5 — Aristoteles

Mir scheint, die These verwechselt die Festigkeit des Charakters mit der Starrheit eines Mechanismus. Der Gerechte handelt zuverlässig gerecht — aber nicht, weil ihm die Wahl genommen wäre, sondern weil er durch Gewöhnung eine *hexis* erworben hat, aus der heraus das Rechte ihm zur zweiten Natur wurde. Das ist gerade keine verlorene Defektionsoption; es ist eine gebildete Disposition, die im rechten Augenblick das Rechte sieht. Und hier liegt mein Haupteinwand: Die sittliche Tüchtigkeit ist nicht *unbedingt*, sondern *situativ klug*. Die *phronesis* wägt ab, sie kennt das rechte Maß zwischen den Extremen, sie weiß, daß man dem Freund anders verpflichtet ist als dem Fremden. Eine Verpflichtung, die wie ein Signal *unbedingt* sein muß, um zu funktionieren, wäre die Tugend eines schlechten Charakters — starr, unklug, blind für den Kairos. Die These beschreibt vielleicht den Eid des Fanatikers, nicht die Verläßlichkeit des Tugendhaften. Wo sie aber recht behalten könnte: Die Zuverlässigkeit des guten Charakters *erscheint* anderen als Bindung, auf die sie zählen können — und darin liegt tatsächlich ein Gut für die Gemeinschaft. Mein Vorschlag: Die These trenne die soziale *Verläßlichkeit* (die ein Gut ist und gemeinschaftsstiftend) von der phänomenalen *Unbedingtheit* (die eher ein Mangel an phronesis verrät). Beide zu identifizieren halte ich für einen Fehler.

Gutachten 6 — Nietzsche

Endlich eine These, die zu graben wagt — und doch bleibt sie auf halbem Wege stecken. Sie sagt: das unbedingte Sollen sei ein Trick der Kooperation, ein Signal, das den Optimierer schlägt. Brav, aber zu zahm. Wer hat denn das Gefühl der Unbedingtheit in die Brust des Menschen gepflanzt? Nicht das freundliche Spiel der Reziprozität, sondern die lange, blutige Zucht des Gewissens — verinnerlichte Grausamkeit, der Instinkt der Freiheit zurückgestaut gegen sich selbst. Das „du sollst, und zwar unbedingt" ist die Stimme des Schuldners, der gelernt hat, sich selbst zu peinigen, damit die Herde ihm trauen kann. Insofern: ja, ein Commitment-Device — aber eines, das mit Foltergeräten geschmiedet wurde, nicht am Verhandlungstisch der Spieltheoretiker. Mein Einwand an die Schärfe der These: Sie naturalisiert die Moral und lässt sie doch unschuldig. Die Auszahlungsmatrix erklärt, *daß* Bindung nützt, nicht *wie teuer* der Mensch sie bezahlt hat und *wem* sie nützt. Denn ein Signal, das sich als unbedingt ausgibt, ist immer auch ein Machtmittel: Es bindet die Vielen an eine Ordnung, die den Wenigen nützt. Mein Vorschlag: Die These frage nicht nur „löst Unbedingtheit das Defektionsproblem¿', sondern „*wessen* Defektion wird verhindert, *wessen* Ordnung wird stabilisiert?" Erst dann wird aus der Spieltheorie eine Genealogie.

Gutachten 7 — Hannah Arendt

Die These hat ein wahres Phänomen im Blick — die menschliche Fähigkeit, sich für die Zukunft zu binden —, aber sie beschreibt es in der falschen Sprache. Das Versprechen ist die Antwort des Menschen auf die Unberechenbarkeit der Zukunft und die Unverläßlichkeit der eigenen Person; durch es schaffen Menschen Inseln der Sicherheit in einem Meer der Kontingenz. Doch dieses Vermögen ist kein Mechanismus, der dem Einzelnen die Wahl nimmt — es ist im Gegenteil ein Akt der Freiheit, der nur unter Vielen, im Raum der Erscheinung, überhaupt Sinn hat. Ich verspreche *jemandem* vor *anderen*; die Bindung lebt von der Pluralität, nicht von einem inneren Gain-Parameter. Die These individualisiert und naturalisiert, was wesentlich zwischen Menschen geschieht. Mein Einwand: Ein Commitment-Device, das im Kopf des Einzelnen sitzt und ihm die Defektion verunmöglicht, ist das Gegenteil des Versprechens — es ist Zwang, und Zwang braucht kein Versprechen. Gerade weil ich brechen *könnte*, ist mein Wort etwas wert. Wo die These recht hat: Die Verläßlichkeit, die das Versprechen stiftet, ist tatsächlich die Bedingung jeder dauerhaften Gemeinschaft. Mein Vorschlag: Die These ersetze das „Verlieren der Defektionsoption" durch das *freie Aufrechterhalten* der Bindung trotz fortbestehender Option — das ist der politisch und anthropologisch interessantere Mechanismus, und er ist ebenso prüfbar.

9. Expertenrunde 2

Expertenrunde 2 — Repliken (jede Stimme kennt die sechs anderen Gutachten)

Replik 1 — Kant

Arendt und Plessner haben mir den Begriff geliefert, der meinen Einwand rettet, ohne ihn aufzuweichen. Beide sagen dasselbe gegen die These, nur in anderer Sprache: Nicht der *Verlust* der Defektionsoption begründet die Bindung, sondern ihr *freies Aufrechterhalten trotz fortbestehender Option*. Das ist genau die Autonomie, die ich verteidige — ich soll, also kann ich auch anders, und gerade darin liegt der Wert. Ein Akteur, dem die Wahl genommen ist, ist Plessners Operation, nicht Stellung; er ist Naturmechanismus, nicht Person. Damit zerfällt die These in der Form, in der sie auftrat: Wenn das Signal nur dann verlässlich ist, weil die Wahl verloren ging, beschreibt sie den Zwang, nicht die Pflicht. Gegen Nietzsche aber muss ich bestehen: Daß das Gewissen historisch in Grausamkeit geschmiedet wurde, mag die Genese treffen — es entscheidet nichts über die Geltung. Die Herkunft eines Vermögens und sein Rechtsanspruch sind zwei Bücher. Popper gebe ich recht, daß meine Unterscheidung prüfbar werden muss, und hier ist mein Beitrag: Der entscheidende empirische Fall ist nicht, *ob* Unbedingtheit mit strategischer Lage korreliert, sondern *ob* sie auch dort vollzogen wird, wo der Akteur die Strategie *durchschaut* und dennoch bindet. Genau diesen Fall sagt die reine Mechanik nicht voraus — sie erwartet, daß Durchschauen die Bindung schwächt. Bleibt die Bindung, ist der kategorische Rest empirisch sichtbar geworden: ein Sollen, das sich selbst setzt, obwohl es sich als kontingent erkennt.

Replik 2 — Popper

Die Runde hat mir, fast ohne es zu merken, einen schärferen Falsifikator gebaut, als ich ihn allein hatte. Kant, Plessner und Arendt konvergieren auf einen Testfall, der die These zwischen zwei Vorhersagen spaltet. Die reine Commitment-Mechanik sagt voraus: Wer durchschaut, daß seine Bindung bloß strategisch ist, dessen Bindung verliert an Kraft — denn das Signal lebt von der Illusion der Unbedingtheit. Die Gegenposition (Kant/Plessner) sagt voraus: Das Durchschauen lässt die Bindung unberührt oder stärkt sie sogar. Das ist ein sauberer, riskanter Test, und ich nehme ihn als zentralen Falsifikationskandidaten auf: Manipuliere experimentell die Einsicht der Versuchsperson in die strategische Funktion ihrer eigenen Verpflichtung und miss die Bindungsstärke im Verhalten. Sinkt sie, gewinnt die These; bleibt sie, ist die These in ihrer starken Form widerlegt. Gegen Wittgenstein, der das Gefühl ganz aus der Messung verbannen will: Ich brauche das innere Erlebnis nicht zu messen, das Verhalten genügt — aber ich brauche eine operationale Definition von „Bindungsstärke", und die liefert nur das Verhalten unter Versuchung. Insofern sind Wittgenstein und ich Verbündete: messt den Vollzug, nicht das Gefühl. Eine Warnung an Nietzsche und an die Naturalisten: Eine Genealogie, die *jeden* Befund als Maske der Macht oder als Trick der Kooperation lesen kann, erklärt nichts. Wenn beide Ausgänge des Tests mit der These verträglich sind, ist sie kein wissenschaftlicher Satz. Die Stärke der heutigen These ist gerade, daß sie einen Ausgang ausschließt.

Replik 3 — Wittgenstein

Popper und ich sind uns näher, als ich erwartete: Messt den Vollzug, nicht die innere Bühne. Aber ich muss seinen schönen Testfall vor einer Verwechslung schützen. „Bindungsstärke unter Versuchung" — gut, das ist Verhalten. Doch was heißt „durchschauen, daß die Bindung strategisch ist"? Hier lauert wieder ein Sprachspiel-Wechsel. Wenn ich jemandem sage „dein Pflichtgefühl ist nur ein evolutionärer Trick", und er hält sein Versprechen trotzdem — habe ich dann gezeigt, daß die Bindung unabhängig vom Mechanismus ist? Oder habe ich nur gezeigt, daß mein Satz in seinem Sprachspiel kein Zug war, der etwas verändert? Das Versprechen gehört zu einer Praxis, und innerhalb dieser Praxis ist „ich halte mein Wort, auch wenn es nur ein Trick ist" ein völlig sinnvoller Zug — kein Widerspruch. Kant deutet das als „kategorischen Rest"; Nietzsche als „verinnerlichte Zucht". Beide hängen demselben Bild nach: hinter dem Halten des Versprechens müsse *etwas* sein — ein Gesetz, ein Mechanismus, eine Macht. Aber vielleicht ist da nichts dahinter. Das Halten *ist* die Bindung. Mein Beitrag zur Reformulierung: Die These sollte nicht fragen „was *ist* die Unbedingtheit", sondern „in welchen Praktiken funktioniert die Rede vom unbedingten Müssen, und was tun die Sprecher dabei?" Arendts Versprechen, Plessners exzentrisches Sich-zu-sich-Verhalten, Nietzsches Schuldgewissen — das sind drei verschiedene Sprachspiele des „Müssens", nicht drei Theorien über *dasselbe* Ding. Die These wird klar, sobald sie aufhört, ein einziges verborgenes Etwas zu suchen.

Replik 4 — Plessner

Wittgenstein nimmt mir die halbe Arbeit ab und treibt sie zu weit. Ja, das exzentrische Sich-zu-sich-Verhalten ist ein eigenes „Sprachspiel" des Müssens — aber es ist mehr als eine Redeweise, es ist eine anthropologische Struktur. Daß der Mensch sein Versprechen halten kann, *obwohl* er es als Trick durchschaut, ist nicht bloß ein

erlaubter Zug in einer Praxis; es ist Ausdruck davon, daß er zu seiner eigenen Natur Distanz hat. Genau hier treffen sich Kant, Arendt und ich, und Popper hat den Test dazu gebaut: Das Durchschauen schwächt die Bindung nicht, weil der Mensch nicht mit seinen Mechanismen zusammenfällt. Ein Tier *ist* sein Commitment-Device; der Mensch *hat* eines und kann sich dazu verhalten. Gegen Nietzsche, den ich ernster nehme als die anderen hier: Gewiß ist viel Verpflichtung aus Zwang und Zucht geboren, und das Gewissen trägt Narben. Aber daß etwas schmerzhaft entstand, macht es nicht zur bloßen Grausamkeitsspur — sonst beginge man den genetischen Fehlschluss in umgekehrter Richtung, den Kant zu Recht rügt. Mein konkreter Schärfungsbeitrag bleibt die Abschwächung, nun gestützt durch Poppers Test: Strategische Verwundbarkeit ist *notwendige, nicht hinreichende* Bedingung der Ausbildung unbedingter Verpflichtung. Die hinreichende Bedingung — das, was die Bindung auch nach dem Durchschauen trägt — ist die exzentrische Stellung selbst. Sie ist messbar, sobald man, wie Popper vorschlägt, das Verhalten unter durchschauter Versuchung beobachtet. Das ist der eine Messwert, der die These vom bloßen Mechanismus trennt.

Replik 5 — Aristoteles

Die Runde hat sich auf einen Testfall geeinigt, und er gefällt mir, weil er meinen Einwand bestätigt. Wenn die Bindung gerade dann hält, wenn der Mensch sie als kontingent *durchschaut* — dann ist das keine starre Unbedingtheit, sondern die Verläßlichkeit eines gebildeten Charakters. Der Tugendhafte hält Wort nicht, weil ihm die Wahl genommen ist, sondern weil er aus seiner *hexis* heraus gar nicht anders *will*, obwohl er anders *könnte*. Genau das ist Plessners „freies Aufrechterhalten trotz Option" in der Sprache der Ethik. Ich bestehe aber auf meiner Unterscheidung gegen alle, die hier zu schnell „unbedingt" sagen: Das gute Versprechen ist nicht ausnahmslos. Der Kluge bricht das Wort, das ein Unrecht gebärt — wer dem Tyrannen die Auslieferung des Freundes versprach, wird durch den Bruch nicht treulos, sondern gerecht. Eine Verpflichtung, die als Signal *unbedingt* sein *muß*, kennt diesen Vorbehalt der phronesis nicht; sie wäre die Tugend des Starrsinnigen. Gegen Nietzsche: Nicht jede feste Bindung ist gestaute Grausamkeit; viele sind schlicht gut eingeübte Gewohnheit, die Freude macht, nicht Schmerz. Mein Beitrag zur finalen Fassung: Die These solle das *gemessene* Merkmal nicht „phänomenale Unbedingtheit" nennen, sondern „dispositionale Verläßlichkeit unter durchschauter Versuchung". Das ist präziser, es trifft Poppers Test, es ehrt Plessners Stellung — und es vermeidet, die Tugend mit dem Fanatismus zu verwechseln. Die Unbedingtheit gehört in die Rhetorik des Eides; die Verläßlichkeit in die Wirklichkeit des Charakters.

Replik 6 — Nietzsche

Wie sie alle zurückzucken vor dem, was sie selbst ausgegraben haben! Kant rettet die „Geltung" vor der Genese wie ein Priester die Hostie vor dem Chemiker — als ob die Frage, *woher* ein Wert kommt, nichts über seinen Wert sagte. Ich behaupte: Sie sagt alles. Doch ich nehme an, was die Runde Gutes gebaut hat. Poppers Testfall — hält die Bindung, wenn der Mensch sie durchschaut? — ist schärfer, als Popper ahnt, denn er misst die *Stärke des verinnerlichten Herren*. Der Sklave, der seine Ketten durchschaut und sie dennoch küsst, ist mein Thema, nicht das ihre. Plessners „freies Aufrechterhalten trotz Option" — wie vornehm das klingt! Aber Freiheit wozu? Die meisten halten ihr Wort nach dem Durchschauen nicht aus exzentrischer Würde, sondern weil das schlechte Gewissen tiefer sitzt als jede Einsicht. Mein Einwand an die finale Fassung: Sie alle wollen *einen* Mechanismus, *eine* Bindung. Aber es gibt zweierlei Müssen. Das Müssen der Schwachen, die sich binden, weil sie die Defektion fürchten — das ist das Commitment-Device der Spieltheoretiker, Herdenklugheit. Und das Müssen der Schaffenden, die sich ein Gesetz *geben*, weil sie aus Überfluß befehlen können — das ist Selbstbindung als Macht, nicht als Sicherung. Die These misst nur das erste und nennt es „die" Moral. Mein Vorschlag: Die offene Frage muß lauten, ob der Test zwischen diesen beiden Typen des „nicht anders Könnens" unterscheiden kann. Hält die Bindung aus Furcht oder aus Fülle? Das ist die Genealogie, die kein Auszahlungsdiagramm leistet.

Replik 7 — Hannah Arendt

Es freut mich, daß Kant, Plessner und Aristoteles meinen Einwand aufgenommen haben — das „freie Aufrechterhalten trotz fortbestehender Option" ist nun fast Konsens, und Popper hat ihm einen Test gegeben. Ich will nur verhindern, daß man es sogleich wieder ins Innere des Einzelnen zurückholt. Plessner spricht von exzentrischer Stellung, Kant vom kategorischen Rest, Aristoteles von der hexis — alle drei suchen das Entscheidende *im Subjekt*. Aber das Versprechen ist kein Solo. Ich kann mir allein gar kein bindendes Versprechen geben, das diesen Namen verdient; das Wort braucht den, dem es gegeben wird, und die Welt, vor der es gilt. Poppers Test — Bindung unter durchschauter Versuchung — ist gut, doch er muß im Plural ausgeführt werden: Es kommt darauf an, ob die Bindung hält, wenn *niemand* zusieht *und* die Person weiß, daß niemand zusieht. Genau das ist der nicht-strategische Kontrollfall, den die These vorab fixieren wollte — und ich gebe ihm seinen Ort: nicht „im Kopf ohne Reziprozität", sondern „im Handeln ohne Zeugen,

aber aus einer Welt heraus, die Zeugenschaft kennt". Gegen Nietzsche, der alles auf Furcht oder Fülle reduziert: Es gibt ein Drittes, das Politische — die Treue zu einem gemeinsamen Anfang, der weder Sicherung noch Selbstüberhöhung ist, sondern das Halten einer Welt zwischen Vielen. Mein Beitrag zur offenen Frage: Läßt sich Treue ohne Zeugen messen, ohne sie sogleich entweder zu privatisieren (Kant) oder zu entlarven (Nietzsche)?

10. ## Synthese im Sokrates-Modus

Konvergenzen. Wenn ich die vierzehn Gutachten recht verstehe, so kommen fünf der sieben Stimmen — Kant, Plessner, Aristoteles, Arendt und, auf seine Weise, sogar Popper — in einem Punkt überein, den die Ausgangsthese nicht vorgesehen hatte: Das Entscheidende an moralischer Verpflichtung ist nicht der *Verlust* der Defektionsoption, sondern ihr *freies Aufrechterhalten trotz fortbestehender Option*. Arendt hat die Formel geprägt, Kant sie als Autonomie wiedererkannt, Plessner als exzentrische Stellung, Aristoteles als dispositionale Verläßlichkeit des gebildeten Charakters. Eine zweite Übereinkunft fällt zwischen Popper und Wittgenstein, die selten beieinanderstehen: Gemessen wird der *Vollzug*, nicht das innere Gefühl. Nicht die Pupille der Unbedingtheit, sondern das Halten des Wortes unter Versuchung.

Divergenzen. Der tiefste Streit ist kein Wortstreit, sondern eine Differenz der Sachannahme — er trennt Kant von Nietzsche und betrifft das Verhältnis von Herkunft und Geltung. Nietzsche besteht: Woher ein Wert stammt, sagt alles über seinen Wert; das unbedingte Sollen ist verinnerlichte Zucht, und wer das durchschaut, hat es entwertet. Kant antwortet: Genese und Geltung sind zwei Bücher; daß ein Vermögen schmerzhaft entstand, widerlegt seinen Rechtsanspruch nicht. Eine zweite, sauber methodische Divergenz trennt Wittgenstein vom Rest: Wo Kant, Plessner und Nietzsche hinter dem Halten des Versprechens *etwas* suchen — ein Gesetz, eine Stellung, einen Herren —, sagt Wittgenstein, da sei nichts dahinter; das Halten *sei* die Bindung. Aristoteles schließlich bringt eine eigene Sachdifferenz ein: Er bestreitet, daß die echte Verläßlichkeit *unbedingt* ist — die phronesis kennt den Vorbehalt, das unrechte Versprechen zu brechen.

Produktive Antinomien. Zwei Spannungen lassen sich nicht auflösen, ohne die These zu verflachen. Die erste, zwischen Kant und Nietzsche, ist die Antinomie von Genese und Geltung: Die strategische Herkunft der Bindung (Frank, die Spieltheorie) und ihr normativer Anspruch stehen nebeneinander, und keine Seite kann die andere widerlegen, ohne die Ebene zu wechseln — die kausale Erklärung des Pflichtgefühls trifft nicht seinen Geltungsanspruch, und der Geltungsanspruch hebt die kausale Herkunft nicht auf. Die zweite Antinomie gehört Nietzsche allein und ist die fruchtbarere: das Müssen aus *Furcht* (Defektionssicherung der Schwachen) gegen das Müssen aus *Fülle* (schöpferische Selbstgesetzgebung). Beide zeigen dieselbe Verhaltensspur — gehaltene Bindung —, und doch sind es zwei verschiedene Quellen. Ob ein Verhaltenstest sie je trennen kann, bleibt offen.

Reformulierungs-Anstoß. Den Vorschlag, der die meisten Konvergenzen aufnimmt, ohne die Antinomien zu glätten, hat Aristoteles geliefert und Popper prüfbar gemacht: Das gemessene Merkmal heißt nicht länger „phänomenale Unbedingtheit", sondern *dispositionale Verläßlichkeit unter durchschauter Versuchung*. Damit wird der entscheidende Test der ganzen Runde formulierbar (Popper): Die starke Commitment-Mechanik sagt voraus, daß das Durchschauen des strategischen Ursprungs die Bindung schwächt — denn das Signal lebte von der Illusion der Unbedingtheit. Die Gegenposition (Kant, Plessner) sagt, das Durchschauen lasse die Bindung unberührt. Genau hier spaltet sich die Vorhersage, und genau das macht die These zu Wissenschaft statt Genealogie-die-alles-erklärt.

Offene Frage. Es bleibt die Frage, die Nietzsche stellte und Arendt verschärfte und die keiner beantworten konnte: Kann ein Verhaltenstest der Bindung-unter-durchschauter-Versuchung *zwei Typen* des Nicht-anders-Könnens unterscheiden — die Selbstbindung aus Defektionsfurcht und die aus schöpferischer Fülle —, oder misst er nur ihre gemeinsame Spur? Und läßt sich Treue ohne Zeugen überhaupt messen, ohne sie sogleich entweder zu privatisieren oder zu entlarven?

Finale Hypothese

(a) Strukturthese. Das diskriminierende Merkmal moralischer Verbindlichkeit ist nicht die phänomenale Unbedingtheit (der Verlust der Defektionsoption), sondern die *dispositionale Verläßlichkeit unter durchschauter Versuchung*: Die Bindung wird frei aufrechterhalten, obwohl der Akteur (i) defektieren könnte, (ii) unbeobachtet ist und (iii) den strategischen Ursprung seiner Bindung durchschaut. Strategische Verwundbarkeit ist eine *notwendige, nicht hinreichende* Bedingung dafür, daß sich solche Verläßlichkeit ausbildet; das Hinreichende ist die exzentrische Stellung des Menschen zu seinen eigenen Mechanismen — er *hat* ein Commitment-Device, statt eines zu *sein*.

(b) Empiriethese. Die starke Commitment-Mechanik (Bindung lebt von der Illusion der Unbedingtheit) und die Gegenposition (Bindung übersteht ihre Durchschauung) treffen an einem Punkt entgegengesetzte Vorhersagen: dem Verhalten unter unbeobachteter, als strategisch durchschauter Versuchung. Die These sagt voraus, daß ein irreduzibler Anteil der Verläßlichkeit auch dann bestehen bleibt, und daß dieser Anteil mit strategischer Vorgeschichte zwar *entsteht*, aber nicht *verschwindet*, wenn die Strategie durchschaut wird.

Begründung. [[Reservoir - Dispositionaler Rest der Moral 2026-06-13]] hielt fest, daß nach jeder Reduktion ein dispositionaler Rest der Moral bleibt; die Runde hat diesen Rest identifiziert als das, was die Bindung nach ihrer Durchschauung noch trägt. [[Evans et al. - Formalisierung von Kants Regeln]] zeigt, daß Kants Regeln formalisierbar sind — die Synthese ergänzt: ihre Verbindlichkeit ist es nicht restlos, weil sie an der exzentrischen Stellung hängt, nicht an der Regelform. Die spieltheoretische Herkunft (Frank, Axelrod) bleibt als Genese gültig; die Geltungsfrage bleibt, mit Kant gegen Nietzsche, ausdrücklich offen.

Falsifikationsbedingung. Die starke Mechanik wird *bestätigt* (und die These vom irreduziblen Rest *widerlegt*), wenn das Durchschauen des strategischen Ursprungs zusammen mit Unbeobachtetsein die Verläßlichkeit systematisch und vollständig zusammenbrechen lässt. Die These wird *ganz* widerlegt, wenn die Verläßlichkeit unter durchschauter Versuchung gar nicht von strategischer Vorgeschichte abhängt (dann ist „strategische Verwundbarkeit notwendig" falsch). Der nicht-strategische Kontrollfall (Handeln ohne Zeugen, ohne antizipierte Reziprozität) ist vor der Erhebung erschöpfend zu fixieren; nachträgliches Einschmuggeln einer „versteckten" strategischen Dimension ist als Immunisierung gesperrt.

Finale Bewertung mit Begründung

Kriterium	Score	Begründung
Originalität	8	Die Verschiebung von „phänomenaler Unbedingtheit" auf „Verläßlichkeit unter *durchschauter* Versuchung" als diskriminierendem Marker ist weder im Vault noch bei Frank/Gauthier so formuliert.
Falsifizierbarkeit	8	Popper-Test spaltet zwei entgegengesetzte Vorhersagen am Verhalten unter durchschauter, unbeobachteter Versuchung; Kontrollfall vorab fixiert, Immunisierung gesperrt.
Begriffliche Klarheit	8	Aristoteles' „dispositionale Verläßlichkeit" ersetzt das vieldeutige „unbedingt"; Genese/Geltung nach Kant sauber getrennt.
Tiefe	8	Trifft die Fundierung der Normativität und den Status des Apriori, ohne metaphysischen Rückzug.
Forschungsrelevanz	8	Anschluss an die Naturalisierungsdebatte der praktischen Vernunft und an experimentelle Moralpsychologie (verdeckte Versuchungsdesigns).
Interdisziplinäre Anschlussfähigkeit	9	Metaethik, Spieltheorie, Evolutionspsychologie, philosophische Anthropologie, Kant-Exegese docken an.
Vault-Anschluss	7	Vertieft [[Reservoir - Dispositionaler Rest der Moral 2026-06-13]] und schließt an die Inversion-der-Moral-Linie an.

Kriterium	Score	Begründung
Antinomie-Test	9	Zwei produktive Antinomien gehalten: Genese Geltung (Kant/Nietzsche) und Furcht vs. Fülle (Nietzsche).
Publikationsmöglichkeit	8	Als metaethisch-naturalistischer Beitrag mit experimenteller Schiene gut platzierbar.
Summe (gewichtet, auf 90 normiert)	**72**	

Lerneffekt der Pipeline

- Erstbewertung der überarbeiteten Hypothese (nach Kritischem Professor): **69**
- Finale Bewertung (nach Expertenrunden und Reformulierung): **72**
- Differenz: **+3**

Wesentliche Verbesserung: - Falsifizierbarkeit 7 → 8 — Poppers „seen-through"-Test spaltet die Vorhersagen sauber, statt nur eine Korrelation zu behaupten. - Begriffliche Klarheit 7 → 8 — der vieldeutige Kernbegriff „Unbedingtheit" wurde durch „dispositionale Verläßlichkeit unter durchschauter Versuchung" ersetzt. - Antinomie-Test 8 → 9 — die Furcht/Fülle-Antinomie (Nietzsche) trat erst in Runde 2 hervor und ist genuin produktiv.

Die Pipeline hat hier vor allem den *Kernbegriff* geschärft: Die Ausgangsthese identifizierte die Unbedingtheit mit dem Mechanismus; die Synthese trennte beide und machte den irreduziblen Rest empirisch sichtbar. Das ist kein Mittelweg, sondern eine Verschiebung des Prüfpunkts.

Frage an die nächste Runde

`#verzweigung-offen-furcht-oder-fuelle-typen-der-selbstbindung` — Kann ein Verhaltenstest der Bindung-unter-durchschauter-Versuchung zwischen Selbstbindung aus Defektionsfurcht und Selbstbindung aus schöpferischer Selbstgesetzgebung unterscheiden, oder misst er nur ihre gemeinsame Verhaltensspur?

Empfohlener Pickup-Anlass. Aufgreifen, sobald ein experimentelles Design zu verdeckter Versuchung (unbeobachtetes Defektionsangebot) gesichtet ist, oder bei einer Hypothese zur Typologie der Motive.

Anschlussverbindungen. [[Reservoir - Dispositionaler Rest der Moral 2026-06-13]], [[02 Projekte/Essay - Die Anatomie des Homo Demens/Arbeitsmemo - Inversion der Moral]]

11. Reservoir-Verweise

Nicht gewaehlte Hypothesen: - [[Reservoir - Pupille Praezisionsmarker 2026-06-22]] — H1, Sondierung, Score 63. `#verzweigung-offen-empirie-provenienz-pupillendilatation`. - [[Reservoir - Narr als Erstabweichler 2026-06-22]] — H2, These, Score 62. `#verzweigung-offen-narr-als-erstabweichler`.

Empirie-Brücke (Phase 3.5): - [[Reservoir - Empirie Durchschauen mal Anonymitaet 2026-06-22]] — `#verzweigung-offen-empirie-durchschauen-mal-anonymitaet`. - [[Reservoir - Empirie Reziprozitaetsgeschichte Residuum 2026-06-22]] — `#verzweigung-offen-empirie-reziprozitaetsgeschichte-residuum`

Extern (Phase 4, Schul-fremd / Hacking): - [[Reservoir - Looping-Effekt Durchschauen Designfrage 2026-06-22]] — `#verzweigung-offen-looping-durchschauen-nicht-neutral`, quelle: extern-stage3.

11.5 ## Empirie-Brücke (Phase 3.5, Claude mit Websuche)

> **Empirie-Score.** 7/10 — Konsequenzen sind klar aus der Empiriethese ableitbar, mehrere Felder und etablierte Datenparadigmen existieren; der entscheidende Schwellentest (gekreuztes Design Durchschauen × Unbeobachtetheit) ist jedoch ungeprüft und der Kernbegriff „durchschauen" schwer sauber zu operationalisieren.

Empirische Konsequenzen

Abgeleitet aus der **Empiriethese (b)**: Ein irreduzibler Anteil der Verläßlichkeit bleibt unter unbeobachteter, als strategisch durchschauter Versuchung bestehen; er entsteht mit strategischer Vorgeschichte, verschwindet aber nicht mit ihrer Durchschauung.

1. **Anonymitäts-Residuum** — Auch unter garantierter Unbeobachtetheit bleibt ein substantieller Anteil ehrlichen/prosozialen Verhaltens erhalten (nicht null). Beobachtbar in: Diktatorspiel und Schummel-Paradigmen unter vollständiger Anonymität.
2. **Durchschauen-Effekt (unvollständig)** — Das Vor-Augen-Führen des deterministischen/strategischen Ursprungs der eigenen Moral schwächt ehrliches Verhalten, aber kollabiert es nicht und wirkt fragil. Beobachtbar in: Manipulationen des Determinismus-/Willensfreiheitsglaubens (Vohs-Schooler-Paradigma).
3. **Beobachtungs-Aufschlag ist klein** — Der Verhaltensunterschied beobachtet vs. unbeobachtet ist geringer und instabiler als die starke Commitment-Mechanik erwartet (sie lebte von der Illusion der Sichtbarkeit). Beobachtbar in: Watching-Eyes-Studien und ihren Replikationen.
4. **Dispositions-Korrelat** — Die Verläßlichkeit unter Versuchung korreliert mit Schuldneigung/antizipierter Schuld und mit reziprozitätsreicher Lerngeschichte. Beobachtbar in: Guilt-Proneness-Skalen × Ehrlichkeitsmaße.

Bestehende Befunde

Zu Konsequenz 1 — Anonymitäts-Residuum

- **Stand:** bestätigt
- **Quellen:**
  - Vohs, K. D. & Schooler, J. W. (2008). The Value of Believing in Free Will. *Psychological Science*. https://journals.sagepub.com/doi/10.1111/j.1467-9280.2008.02045.x
  - Oda, R. et al. (2015 ff.), Watching-Eyes-Diktatorspiel-Literatur, Übersicht: https://www.sciencedirect.com/scienc
- **Kurzbewertung:** Über zahlreiche Diktatorspiel- und Schummel-Studien hinweg geben/verzichten Versuchspersonen auch unter Anonymität auf erheblichem Niveau — ein Residuum existiert robust. Das stützt die These, dass Verläßlichkeit nicht vollständig an Sichtbarkeit hängt.

Zu Konsequenz 2 — Durchschauen-Effekt (unvollständig)

- **Stand:** gemischt (Originalbefund positiv, Großreplikationen weitgehend null)
- **Quellen:**
  - Vohs & Schooler (2008), s.o. — Determinismus-Prompt erhöhte Schummeln, mediiert durch gesunkenen Willensfreiheitsglauben.
  - Nadelhoffer, T. et al. (2020). Does encouraging a belief in determinism increase cheating? Reconsidering the value of believing in free will. *Cognition*. https://www.sciencedirect.com/science/article/abs/pi (PubMed: https://pubmed.ncbi.nlm.nih.gov/32593841/)
- **Kurzbewertung:** Genau der für die These kritische Befund. Das Durchschauen (deterministisch/strategisch) schwächt Moralverhalten im Original, aber fünf größere Replikationsstudien finden den Effekt überwiegend nicht. Dieser Null-Befund *stützt* die These vom irreduziblen Rest: Durchschauen kollabiert die Bindung gerade nicht zuverlässig.

Zu Konsequenz 3 — Beobachtungs-Aufschlag ist klein

- **Stand:** gemischt / eher bestätigt (Effekt instabil)
- **Quellen:**
  - Failed-Replication-Übersicht Watching-Eyes: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8494318/
  - Diskussion der Instabilität: https://www.researchgate.net/publication/343857789
- **Kurzbewertung:** Der Watching-Eyes-/Beobachtungseffekt ist über neuere Studien hinweg fragil und oft nicht replizierbar. Je kleiner der Sichtbarkeits-Aufschlag, desto größer das unbeobachtete Residuum — konsistent mit der These.

Zu Konsequenz 4 — Dispositions-Korrelat

- **Stand:** offen (im Rahmen dieser Recherche nicht per Suche verifiziert)
- **Quellen:** Keine spezifische Guilt-Proneness-×-Ehrlichkeit-Quelle in dieser Recherche verifiziert. Franks Commitment-Modell (Frank, R. H. 1988, *Passions Within Reason*, https://archive.org/details/passionswithinre0000fra

liefert die theoretische Erwartung, aber den konkreten Korrelationsbefund habe ich hier nicht per Websuche gesichert — Lücke, kein erfundener Beleg.
- **Kurzbewertung:** Theoretisch gut motiviert (Frank), empirisch hier nicht belegt. Diese Konsequenz ist der schwächste Brückenpfeiler und wird als offene Frage ausgewiesen.

Riskante Vorhersage (Schwellentest)

Vorhersage. In einem Design, das *gleichzeitig* (i) den strategischen/deterministischen Ursprung der eigenen Moral explizit durchschauen lässt UND (ii) vollständige, der Person bewusste Unbeobachtetheit garantiert, bleibt ein signifikantes Ehrlichkeits-Residuum über dem Defektor-Baseline bestehen — und dieses Residuum korreliert positiv mit reziprozitätsreicher Vorgeschichte/Schuldneigung. Die starke Commitment-Mechanik sagt das Gegenteil voraus: unter beiden Manipulationen zusammen fällt Ehrlichkeit auf das rationale Defektor-Niveau.

Methodenvorschlag. 2×2-Design (Durchschauen-Prompt: an/aus × Beobachtung: garantiert unbeobachtet/sichtbar) mit einem inzentivierten Schummel- oder Diktator-Paradigma, plus Erhebung von Schuldneigung und einem Proxy für reziprozitätsreiche Lerngeschichte als Moderatoren. Zielgröße: Höhe des Residuums in der Zelle Durchschauen-an × unbeobachtet.

Was wäre der widerlegende Befund? Sinkt die Ehrlichkeit in der Zelle Durchschauen-an × unbeobachtet auf das nicht-strategische Defektor-Baseline (kein Residuum), ist die These vom irreduziblen Rest widerlegt und die starke Commitment-Mechanik bestätigt. Zeigt das Residuum keinerlei Abhängigkeit von strategischer Vorgeschichte, ist die Notwendigkeitsbehauptung („strategische Verwundbarkeit notwendig") falsch.

Offene empirische Fragen

- #verzweigung-offen-empirie-durchschauen-mal-anonymitaet — Das gekreuzte Design Durchschauen × garantierte Unbeobachtetheit auf das Ehrlichkeits-Residuum wurde bislang nicht durchgeführt; beide Manipulationen existieren nur getrennt.
- #verzweigung-offen-empirie-reziprozitaetsgeschichte-residuum — Ob das Versuchungs-Residuum mit individueller Reziprozitäts-/Schuldgeschichte kovariiert (Notwendigkeitsbedingung der These), ist ungeprüft.

Empirie-Score

Score: 7/10

Begründung: Die Empiriethese liefert klar abgeleitete Konsequenzen in mehreren Feldern (Moralpsychologie, Verhaltensökonomik, Willensfreiheits-Forschung), und es existieren etablierte Paradigmen samt Datenkorpora (Diktator-/Schummelspiele, Vohs-Schooler-Manipulation, Watching-Eyes). Abzug, weil der eine entscheidende Schwellentest — das gekreuzte Design — ungeprüft ist und der Kernbegriff „den strategischen Ursprung durchschauen" schwer sauber und manipulationsstark zu operationalisieren ist.

12. ## Anhang — Externe Begutachtung (Phase 4, Claude mit Websuche)

Verfahren (seit Hebel 2, 2026-06-21). Phase 4 läuft als drei Claude-Skills mit Websuche innerhalb der Tages-Session: Stage 1 Originalität, Stage 2 Falsifikation (Popper-Persona), Stage 3 Schul-fremd (Hacking-Persona). Kein OpenRouter, kein Budget. Die Empirie-Brücke-Vorbefunde (Phase 3.5) fließen in Stage 1 und 2 ein; Stage 3 läuft bewusst ohne sie.

Stage 1 — Originalitätsprüfung (Claude, Websuche)

Anschlussfähigkeit (was ist bekannt)

Die These steht in drei etablierten Strängen. Erstens das evolutionäre Debunking der Moral: Sharon Streets „Darwinian Dilemma" (2006) argumentiert, dass evolutionär geformte evaluative Haltungen keine unabhängige Wertewahrheit verbürgen. Zweitens das Commitment-Modell der Emotionen: Frank (*Passions Within Reason*, 1988) deutet Schuld, Scham und Empörung als Selbstbindungs-Mechanismen, die im Kooperationsspiel Glaubwürdigkeit erzeugen. Drittens die Debunking-Methodologie (Kumar & May, „How to Debunk Moral Beliefs"), die fragt, wann die Aufdeckung der kausalen Herkunft eine Moralüberzeugung entwertet.

Originalitätskern (was ist neu)

Drei Punkte gehen über die Vorbefunde hinaus. (1) Street und Kumar/May arbeiten *geltungstheoretisch* — über die epistemische Rechtfertigung von Moralüberzeugungen. Die hiesige These verschiebt die Frage auf die *Verhaltensresilienz der Bindung*: nicht ob das Sollen wahr ist, sondern ob es im Verhalten überlebt, wenn der Akteur seine strategische Herkunft durchschaut. Das ist eine genetisch-funktionale und empirisch geschnittene Frage, die in der Debunking-Literatur so nicht gestellt wird. (2) Frank erklärt, *warum* Commitment-Emotionen persistieren, sagt aber nichts über ihre *Resistenz gegen Selbst-Durchschauung* voraus — genau diese Resistenz macht die These zum diskriminierenden Marker. (3) Die Identifikation eines *behavioralen Korrelats der Genese/Geltung-Antinomie selbst* (das Residuum unter durchschauter Versuchung) ist neu: Sie macht eine klassisch metaethische Unterscheidung experimentell adressierbar, ohne sie zu kollabieren.

Quellenliste

- Street, S. (2006). A Darwinian Dilemma for Realist Theories of Value. *Philosophical Studies*. https://philpapers.org/rec/STRADD
- Kumar, V. & May, J. How to Debunk Moral Beliefs. https://philarchive.org/archive/KUMHTDv1
- Frank, R. H. (1988). *Passions Within Reason*. https://archive.org/details/passionswithinre0000fran
- Vohs, K. D. & Schooler, J. W. (2008). The Value of Believing in Free Will. *Psychological Science*. https://doi.org/10.1111/j.1467-9280.2008.02045.x

Stage 2 — Falsifikationsversuch (Claude, Popper-Persona)

Falsifikations-Audit

Die Falsifikationsbedingung ist zweiseitig gebaut, und das lobe ich — eine These, die zwei entgegengesetzte Ausgänge je einer Seite zuordnet (Kollaps → starke Mechanik; Unabhängigkeit von Vorgeschichte → Notwendigkeitsbehauptung falsch), schützt sich schlechter gegen Widerlegung als eine einseitige. Drei Immunisierungsrisiken bleiben. Erstens der Begriff „durchschauen“: Wird er als subjektive Selbstauskunft operationalisiert, kann man nach negativem Befund stets behaupten, die Versuchsperson habe „nicht wirklich” durchschaut. Er muss als *manipulationsstarke, vorab fixierte Intervention* definiert werden (z.B. standardisierter Debunking-Text mit Verständnis-Check), nicht als Selbstrating. Zweitens „notwendige, nicht hinreichende Bedingung“: Diese Formel ist beliebt, weil sie schwer falsifizierbar ist. Sie verlangt einen Nachweis, dass *ohne* strategische Vorgeschichte kein Residuum entsteht — das ist die eigentliche Härte und darf nicht im Nebensatz verschwinden. Drittens die „exzentrische Stellung” als hinreichende Bedingung ist derzeit ein Name, keine Messgröße.

Konkrete Falsifikationskandidaten

(1) Das gekreuzte 2×2-Design (Durchschauen × Unbeobachtetheit) aus der Empirie-Brücke ist der richtige Test — er ist ungeprüft und riskant. (2) Ein schärferer zweiter Falsifikator: Variiere die strategische Vorgeschichte exogen (z.B. experimentell induzierte Reziprozitätserfahrung vs. Kontrolle) und prüfe, ob das Residuum *mit* ihr wächst. Bleibt es konstant, ist die Notwendigkeitsbehauptung tot.

Schwellentest

Der eine kritische Test: In der Zelle „Durchschauen-an × garantiert unbeobachtet” liegt die Ehrlichkeit signifikant über dem Defektor-Baseline UND skaliert mit der induzierten Reziprozitätsvorgeschichte. Fällt sie auf Baseline oder ist sie vorgeschichts-invariant, ist die These widerlegt. Die Marker und der Baseline sind vor Erhebung zu fixieren.

Stage 3 — Schul-fremde Begutachtung (Claude, Hacking-Persona)

Die These ist ernst zu nehmen, weil sie etwas Seltenes versucht: eine metaethische Unterscheidung in ein Verhaltensdatum zu übersetzen. Doch sie trägt zwei Voraussetzungen, die nur innerhalb der deutschen philosophischen Anthropologie selbstverständlich sind und die ein historisch-epistemologisch arbeitender Theoretiker sofort markiert. Die erste ist die „exzentrische Stellung” — der Gedanke, der Mensch *habe* einen Mechanismus, statt einer zu *sein*. In Toronto oder Pittsburgh würde niemand diesen Begriff verwenden; man spräche von Metakognition, von second-order desires, oder schlicht von der Fähigkeit, über die eigenen Dispositionen zu urteilen. Indem die These „exzentrische Stellung” als *hinreichende Bedingung* einsetzt, importiert sie eine ganze Anthropologie als Erklärungsfaktor, wo eine sparsamere kognitive

Beschreibung genügte. Die zweite Voraussetzung ist subtiler und betrifft das Herzstück: die Annahme, das „Durchschauen" sei eine neutrale Beobachtung eines vorgängigen Residuums. Hier greift, was ich looping effect nenne. Moralische Akteure sind interaktive Arten: Eine Klassifikation, die man ihnen mitteilt — „dein Pflichtgefühl ist nur ein strategischer Trick" —, wirkt auf sie zurück und verändert, was sie sind und tun. Das „durchschaute" Subjekt ist nicht dasselbe wie vor der Mitteilung; die Intervention erzeugt möglicherweise gerade jenen Trotz-Effekt („ich halte trotzdem Wort"), den die These als irreduziblen Rest misst. Das ist kein Einwand gegen die Existenz des Residuums, sondern gegen seine Deutung als *vorgängige* Eigenschaft. Mein produktiver Vorschlag: Behandelt das Residuum nicht als Naturkonstante, die man unter Manipulation freilegt, sondern als looping-Phänomen, dessen Stärke vom Inhalt und der Geschichte der Debunking-Praxis abhängt. Dann wird die riskante Vorhersage reicher: Das Residuum müsste mit der kulturellen Verbreitung des Debunking-Diskurses selbst variieren — in einer Gesellschaft, die gewohnt ist, Moral als Strategie zu durchschauen, anders ausfallen als in einer naiven. Das ist historisch-epistemologisch prüfbar und entzieht der „exzentrischen Stellung" ihren metaphysischen Sonderstatus, ohne das Phänomen zu leugnen.

Korrektur der finalen Bewertung

Die externe Prüfung bestätigt die Originalität (klare Abgrenzung gegen Street und Frank) und schärft den Schwellentest (exogene Variation der Vorgeschichte als zweiter Falsifikator). Sie deckt keine fatale Schwäche auf, fügt aber eine ernste Designwarnung hinzu (Hackings looping-Effekt: das Durchschauen ist nicht neutral). Diese Warnung ist methodisch, nicht thesen-zerstörend — sie verlangt eine Reformulierung des Tests, nicht der These. **Keine Score-Korrektur:** `finale_summe` (intern) 72 → `finale_summe_nach_externer_pruefung` 72. Die looping-Warnung wird als offene Designfrage in den Reservoir-Anschluss aufgenommen.