Inhaltsverzeichnis

Hypothesentag-Gutachten

Die Gewinnerthese

Vergebung als Fehlerkorrektur kooperativer Gleichgewichte

Initiale Fassung

2. Verzweigungs-Pickup

- **warm_pick:** `empirie-reaktion-zeichen` (Makro-Thema Erkenntnistheorie_Cassirer, erfasst 2026-06-15) → Material für **H1**.
- **cold_pick (Diversitäts-Pflicht):** Virgin Node `Kriz - Systemtheorie als Strukturwissenschaft` (Makro-Thema Systemtheorie_Luhmann, noch nie verarbeitet) → Vault-Anker für **H2**. Generation-Directive: H2 im Kriz-Vokabular halten, nicht ins Cassirer/Friston-Zentrum übersetzen.
- **zieldomaene (H2-Disziplin):** Anthropologie (Plessner, Gehlen, Scheler, Tomasello, Donald, Deacon).
- **exploration_domaine (H3):** Spieltheorie & Kooperationsforschung (Nash, Axelrod, Schelling, Skyrms) → Material für **H3**.
- **devil_advocate:** nicht aktiv (Montag; keine Drift; kein Echo-Chamber-Alarm).
- **schwerpunkt_woche:** ethisch_praktische_hypothese (Methoden-Typ empirisch_pruefbar, mit normativ-begriffsanalytischem Kern).

Reformulierte Fassung (nach Kritischem Professor)

Hypothese 1 (warm_pick: empirie-reaktion-zeichen) — Bedeutung als Sequenz-Selbstbezug

Kernsatz. Die bedeutungstragende Relation, die den Übergang von Ausdruck zu Darstellung in der paläolithischen Zeichenpraxis antreibt, liegt nicht zwischen zwei Köpfen (Sender und Empfänger), sondern

in der Reaktion späterer Zeichen auf frühere auf derselben Fläche: Ein Zeichen stabilisiert sich, indem es ein vorausgegangenes Zeichen beantwortet. Bedeutung ist sequenzieller Anschluss, nicht dyadische Übertragung.

Begründung. Der Strang `empirie-reaktion-zeichen` (aus [[06 Hypothesentag/2026-06-15]]) hat als offenen Test hinterlassen, ob in archäologischen Sequenzen spätere Markierungen messbar auf frühere reagieren — das war die Luhmann-seitige Probe der Signalgleichgewichts-These (Stabilisierung *in* der Folge, nicht *zwischen* zwei Köpfen). Cassirers Beschreibung der Höhlenmalerei als erste symbolische Form ([[Cassirer - Höhlenmalerei und frühe symbolische Form]]) lässt offen, ob die symbolische Funktion zuerst als Mitteilung an einen Anderen oder als Antwort auf ein eigenes früheres Zeichen entsteht. Die These verschiebt den Ursprungsort: Das früheste Symbolische ist ein Selbstgespräch der Fläche, kein Dialog der Personen. Damit wird der Übergang Ausdruck→Darstellung an einer beobachtbaren Stelle lokalisiert — der formalen und räumlichen Abhängigkeit eines Zeichens von einem vorausgehenden.

Falsifikationsbedingung. Widerlegt, wenn die Sequenzanalyse mehrschichtiger Panels (Übermalungen, Hinzufügungen über Schichtgrenzen) zeigt, dass spätere Markierungen formal und positional statistisch unabhängig von früheren sind, sobald die physischen Flächenrestriktionen kontrolliert sind — dann müsste Bedeutung doch dyadisch (zwischen Personen) verortet werden.

Quelle. warm_pick — verfolgt den Strang `empirie-reaktion-zeichen` aus [[06 Hypothesentag/2026-06-15]].

Expertenrunden und Synthese

7. Bewertung mit dreifacher Klassifikation

Klassifikation pro Hypothese

Hypothese 1 — Bedeutung als Sequenz-Selbstbezug

- **Themenfeld:** epistemologisch_systemtheoretisch (Anschlussfähigkeit, symbolische Funktion, Sequenz-Selbstbezug).
- **Methoden-Typ:** empirisch_pruefbar (mit begriffsanalytischem Kern) — operationalisierte Sequenz-Abhängigkeit; Falsifizierbarkeit-Anker max 10. **Pflichttest Operationalisierung:** bestanden, aber datierungsabhängig (Diskriminanzkriterium definiert).
- **Reichweiten-Klasse:** these — schließt eine offene Verzweigung, verbindet 1–2 dichte Knoten, eigenständig falsifizierbar.

Hypothese 2 — Der Tod als systemisch verarbeiteter Strukturverlust

- **Themenfeld:** epistemologisch_systemtheoretisch (Systemverhalten: Elementverlust, Restabilisierung; Heuristik „bei systemtheoretisch *und* anthropologisch → systemtheoretisch").
- **Methoden-Typ:** empirisch_pruefbar mit begriffsanalytisch-methodologischem Kern (Umdefinition des Schwellenprädikats von individuell zu relational). **Pflichttest Operationalisierung:** bedingt bestanden — die Empiriethese hängt an der Verfügbarkeit einer Konstellation, in der die Markerklassen auseinanderlaufen.
- **Reichweiten-Klasse:** these, **forschungsprogramm_kandidat: true** (öffnet eine Architektur: Schwelle als strukturelles statt kognitives Prädikat; mehrere Linien — Bestattung, Rollenneubesetzung, Dominanzasymmetrie).

Hypothese 3 — Vergebung als Fehlerkorrektur kooperativer Gleichgewichte

- **Themenfeld:** ethisch_praktische_hypothese (Vergebung, Verzeihen, Tugend; Arendt/Nietzsche/Aristoteles).
- **Methoden-Typ:** empirisch_pruefbar (mit begriffsanalytisch-normativer Trennung Verhaltens-/Anerkennungskomponente). **Pflichttest Operationalisierung:** bestanden — Vergebungsrate × Rauschniveau, ceteris-paribus ergänzt; Falsifizierbarkeit-Anker max 10.
- **Reichweiten-Klasse:** these — eigenständig publizierbar, klare Falsifikation, verbindet Vault-Ethikstrang (Arendt) mit externer Domäne.

Pipeline-Regel-Check: Kein Methoden-Typ `methodologisch_wissenschaftstheoretisch` heute (0 2). Mindestens eine empirisch-prüfbare These vorhanden (alle drei) .

9-Kriterien-Tabelle (gewichtet, auf 90 normiert)

Kriterium	H1	H2	H3
Originalität (×1.6)	6 — Verfeinerung im etablierten Signalgleichgewichts-Strang; Übermalungssemiotik existiert.	8 — systemische Umverortung der Todesschwelle (System prozessiert Elementverlust), klar abseits des jüngsten Verbindlichkeits-Clusters.	8 — spieltheoretische Lesart der Vergebung als Rausch-Fehlerkorrektur ist im Vault neu; Reframe gegen Arendt original.
Falsifizierbarkeit (×1.0)	8 — Diskriminanzkriterium operationalisiert, aber an relativer Chronologie aufgehängt.	5 — operationalisiert, doch an Verfügbarkeit einer Divergenz-Konstellation gebunden (Immunisierungsrisiko selbst markiert).	9 — sauberer Test (Vergebung × Rauschen), ceteris-paribus ergänzt, experimentell prüfbar.
Begriffliche Klarheit (×1.0)	7 — „Anschluss" klar; „Reaktion" nun definiert, Restunschärfe bleibt.	7 — entmetaphorisiert: „System bemerkt" = beobachtbare Regelhaftigkeit.	8 — Äquivokation behoben (Verhaltens- vs. Anerkennungskomponente getrennt).
Tiefe (×1.2)	6 — berührt Symbol-Ursprung, bleibt lokal-methodisch.	9 — greift in die Anthropogenese und die Konstitution des Sozialen.	7 — Grund der Stabilität moralischer Praxis, aber funktionalistisch.
Forschungsrelevanz (×1.0)	7 — kognitive Archäologie / Parietalkunst-Semiotik.	7 — Social-Brain-Debatte, Mortalitäts-/Bestattungsarchäologie.	9 — Evolution der Kooperation, Verhaltensökonomik, Moralpsychologie — laufende Debatten.
Interdisziplinäre Anschlussfähigkeit (×1.0)	6 — Archäologie, Semiotik, Systemtheorie.	8 — Archäologie, Anthropologie, Soziologie/Systemtheorie, Primatologie.	9 — Spieltheorie, Evolutionsbiologie, Moralphilosophie, Psychologie, Soziologie.
Vault-Anschluss (×1.0)	9 — vertieft dichten Knoten + schließt warm-Verzweigung direkt.	8 — Bestattungsstränge + Kriz-Virgin-Node (Anker dünner).	6 — Anschluss über Arendt/Ethikstrang, aber externe Expedition.
Antinomie-Test (×1.2)	6 — Gegenthese (dyadisch) plausibel, aber Falsch-Dichotomie schwächt Produktivität.	9 — individuell-kognitive Schwelle (Scheler/Heidegger) genauso plausibel; produktive Spannung Individuum/Struktur.	9 — Vergebung als adaptive Korrektur vs. Vergebung als gerade *nicht* kalkulierbarer moralischer Akt (Arendt/Nietzsche).
Publikationsmöglichkeit (×1.0)	6 — erst mit empirischer Studie tragfähig; Nische.	6 — spekulativ, braucht empirische Schiene.	8 — passt in Biology & Philosophy / Ethik-Venues.
Summe (gewichtet, auf 90 normiert)	**59**	**67**	**71**

Ergebnis. Hypothese 3 (*Vergebung als Fehlerkorrektur kooperativer Gleichgewichte*) erreicht mit **71** den höchsten Score und geht in die Expertenrunde. H1 (59) und H2 (67) wandern ins Reservoir. H2 wird als **forschungsprogramm_kandidat** markiert.

Drei verschiedene thematische Räume bestätigt: H1 Semiotik/Sequenz, H2 sozial-systemische Anthropologie, H3 spieltheoretische Ethik. Keine ist ins Vokabular einer anderen übersetzbar.

Reservoir-Profile (nicht gewählte Hypothesen)

H1 → `#reservoir-bedeutung-sequenz-selbstbezug`

- **Score:** 59/90 (Rang 3). Methoden-Typ empirisch_pruefbar. Schwäche: Datierungshypothek, Originalität inkrementell.
- **Pickup-Anlass:** Tag mit kognitionsarchäologischem oder semiotischem Schwerpunkt; gezielte Sequenzanalyse von Übermalungen.
- **Anschluss:** [[Cassirer - Höhlenmalerei und frühe symbolische Form]], [[Reservoir - Empirie Reaktion Zeichen auf Zeichen 2026-06-15]].

H2 → `#reservoir-tod-strukturverlust` (forschungsprogramm_kandidat)

- **Score:** 67/90 (Rang 2). Methoden-Typ empirisch_pruefbar mit begriffsanalytisch-methodologischem Kern. Stärke: Tiefe (9), Antinomie (9). Schwäche: Falsifizierbarkeit (5).
- **Pickup-Anlass:** Tag mit anthropogenetischem oder systemtheoretisch-soziologischem Schwerpunkt; Anschluss an Social-Brain-Hypothese (Dunbar) und Bestattungsarchäologie.
- **Anschluss:** [[Kriz - Systemtheorie als Strukturwissenschaft]], Bestattungsstränge des Hypothesentags.

8. Expertenrunde 1 — unabhängige Gutachten

Panel-Konfiguration. Hypothese als `ethisch_praktische_hypothese` klassifiziert (Vergebung/Verzeihen, Tugend, moralische Praxis). Panel `ethisch_praktische_hypothese`. Sekundär epistemologisch-systemtheoretisch (Gleichgewicht, Defektion), dokumentiert.

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

Gutachten 1 — Kant

Man muss hier die Bedingungen der Möglichkeit sauber scheiden. Dass eine vergebende Disposition unter Rauschen die Kooperation erhält, ist eine Aussage über die *Legalität* des Verhaltens — über seine Übereinstimmung mit einem vorteilhaften Muster. Über die *Moralität* der Vergebung ist damit nichts gesagt. Eine Vergebung, die geübt wird, weil sie sich auszahlt, ist keine Vergebung aus Pflicht, sondern aus Neigung zum eigenen Vorteil — und genau an dieser Stelle droht die These, das Sittliche auf einen Mechanismus zu reduzieren, der seinem Begriff nach heteronom ist. Ich frage nach der Maxime: Wer vergibt, um ein Gleichgewicht zu stabilisieren, handelt nach der Maxime der Klugheit, nicht des Gesetzes. Das ist kein Einwand gegen die empirische Regelmäßigkeit — die räume ich ein —, sondern gegen den Reduktionsanspruch, den die Strukturthese in ihrer ersten Fassung noch trug und jetzt zurücknimmt; gut so. Was bleibt, ist eine Doppelfrage, die die These schärfen würde: Erklärt der Mechanismus, *dass* Wesen, die vergeben können, sich erhalten — oder beansprucht er auch zu erklären, *was* Vergeben als sittlicher Akt ist? Nur das Erste ist ihm zugänglich. Ich empfehle, die These ausdrücklich als Aussage über die Erhaltungsbedingung einer Praxis zu führen, nicht über ihren Geltungsgrund. Dann ist sie verträglich mit der Autonomie des urteilenden Subjekts, das auch dort vergibt, wo es nichts einbringt.

Gutachten 2 — Popper

Methodisch ist die überarbeitete Fassung deutlich besser als die erste — die ceteris-paribus-Klausel für Reputationssichtbarkeit und Beziehungsdauer war überfällig. Aber ich verlange Schärfe an zwei Stellen. Erstens: Ist das Maß für „Rauschen" unabhängig vom Maß für „Vergebung" erhoben? Wenn die Rauschintensität aus demselben Verhaltensprotokoll geschätzt wird, in dem auch die Vergebungsrate steckt, droht ein Konfirmationszirkel — die These bestätigte sich dann selbst. Man braucht eine *exogen* manipulierte Fehlerwahrscheinlichkeit, etwa im kontrollierten Experiment mit eingespielten Wahrnehmungsfehlern. Zweitens: Der eigentliche Schwellentest darf nicht „Vergebung steigt mit Rauschen" lauten — das ist eine Korrelation, die viele Modelle vorhersagen. Riskant wird die These erst, wenn sie sagt: Nachsichtige Strategien schlagen unnachsichtige *spezifisch unter Rauschen und nicht ohne*, und es gibt ein Rauschniveau, oberhalb dessen Unnachsichtigkeit kollabiert. Das ist falsifizierbar und keineswegs trivial — Axelrods und Nowaks Befunde legen es nahe, widerlegen es aber nicht für menschliches Verhalten. Solange dieser eine entscheidende Test benannt und nicht durch Hilfsannahmen aufgeweicht wird, halte ich die These für eine respektable, prüfbare Vermutung. Die Gefahr ist die Immunisierung durch nachträgliche Reputationsargumente — die müssen vorab kontrolliert, nicht hinterher eingeführt werden.

Gutachten 3 — Wittgenstein

Welches Sprachspiel wird hier gespielt? „Vergebung" gehört in das Spiel von Schuld, Reue, Anrede, Loslassen — ein Spiel mit einem Gegenüber, das adressiert wird. „forgiveness" als Parameter in einer Auszahlungsmatrix gehört in ein ganz anderes Spiel: Es ist die Wahrscheinlichkeit, nach einer Defektion wieder zu kooperieren. Generous Tit-for-Tat vergibt niemandem; es kooperiert mit einer gewissen Rate. Die These tut gut daran, die Verhaltens- von der Anerkennungskomponente zu trennen — aber dann muss sie ehrlich fragen, ob das, was sie erklärt, noch *Vergebung* heißt oder ob sie das Thema gewechselt hat. Mein Verdacht: Sie erklärt etwas, das sich wie Vergebung *verhält*, und nennt es so, weil die Verhaltensoberfläche gleich aussieht. Das ist erlaubt, aber es muss markiert werden, sonst tarnt sich eine begriffliche Verschiebung als Sachentdeckung. Die Probe: Würden wir von einem Automaten, der mit Wahrscheinlichkeit p wieder kooperiert, sagen, er habe „vergeben"? Wir würden zögern — und dieses Zögern ist der ganze begriffliche Befund. Ich schlage vor, im Kernsatz nicht „Vergebung ist die Fehlerkorrektur", sondern „das *Vergebungsverhalten* hat die Funktion der Fehlerkorrektur" zu sagen. Der Unterschied ist nicht Pedanterie, sondern die Grenze des Erklärten.

Gutachten 4 — Plessner

Ich bestehe auf einer schwächeren Form Ihrer These. Dass das vergebende Verhalten unter Rauschen adaptiv ist, ist eine notwendige Hintergrundbedingung seiner Dauerhaftigkeit — nicht sein Wesen. Was an der Vergebung als phänomenale Erfahrung bleibt, geht in der Auszahlung nicht auf: das Erleiden des Unrechts, das Innehalten, das Freigeben des Anderen aus der Festlegung auf seine Tat. Der Mensch ist hier bei sich, indem er sich zu seinem eigenen Ressentiment verhält — er kann eine Stellung zu seinem Gekränktsein einnehmen. Genau diese exzentrische Positionalität fehlt im Modell, das nur Reize und Antworten kennt. Ein Tier in einem iterierten Spiel „verzeiht" nicht, es reagiert. Ich bestreite den Mechanismus nicht; ich bestreite, dass er das Phänomen ausschöpft. Die anthropologische Differenz ist hier konstitutiv, nicht funktional: Vergeben setzt voraus, dass einer sich von der Unmittelbarkeit seiner Reaktion distanzieren *kann*. Meine Empfehlung zur Schärfung: Führen Sie die Funktion als Erhaltungsbedingung, und halten Sie daneben ausdrücklich die Stellungnahme-Fähigkeit als das, was erst erklärt, warum Menschen vergeben können und Strategien es nur simulieren. Dann ist die These stark, weil bescheiden.

Gutachten 5 — Aristoteles

Vergebung — syngnome — ist eine eingewohnte Disposition, eine hexis, keine Wahrscheinlichkeit. Wer recht erzogen ist, vergibt dem Richtigen, im richtigen Maß, zur richtigen Zeit; das verlangt phronesis, praktische Klugheit, die situativ entscheidet und sich keiner festen Rate fügt. Hier liegt mein Einwand gegen das Modell: Eine konstante Vergebungswahrscheinlichkeit p ist gerade *nicht* Tugend, sondern ihr Zerrbild — der Tugendhafte handelt nicht nach Regel, sondern nach der mittleren Mitte zwischen zwei Lastern. Zu viel Nachsicht macht zum Fußabtreter, zu wenig zum Rachsüchtigen; die Tugend ist die mesotes dazwischen, und sie ist konkret, nicht abstrakt. Doch ich gestehe der These einen wahren Kern zu: Diese Disposition wird durch Übung erworben, in wiederholtem Tun — und das wiederholte Spiel ist genau der Ort solcher Einübung. Insofern beschreibt das Modell nicht die Tugend, aber ihren Erwerbsweg. Mein Schärfungsvorschlag: Die These sollte nicht eine feste Vergebungsrate vorhersagen, sondern dass die *Bandbreite gelingender Vergebung* — das richtige Maß — sich mit dem Rauschniveau verschiebt. Mehr Fehlbarkeit der Wahrnehmung verlangt mehr Nachsicht als Mitte; weniger weniger. Das verbindet die mesotes mit Ihrem Rauschparameter, ohne die Tugend zur Strategie zu verflachen.

Gutachten 6 — Nietzsche

Ihr fragt, *warum* vergeben wird — ich frage, was die Vergebung verbirgt. Eure „Fehlerkorrektur kooperativer Gleichgewichte" ist die Selbstbeschmeichelung der Herde: Sie macht aus dem Unvermögen zu vergelten eine Klugheit. Wer nicht zurückschlagen *kann*, nennt es großmütig „Nachsicht" und beruft sich auf die Auszahlung. Das ist Ressentiment im Kostüm der Spieltheorie. Aber — und hier wird es interessant — es gibt eine zweite Vergebung, die euer Modell mit der ersten verwechselt: die Vergebung aus Überfluss, die des Starken, der seine Feinde nicht lange ernst nehmen *kann*, weil er zu viel zu schaffen hat. Diese vergibt nicht, um ein Gleichgewicht zu retten, sondern weil das Festhalten unter ihrer Würde ist. Euer „Generous Tit-for-Tat" kann beide nicht unterscheiden — und das ist sein eigentlicher Defekt. Eine Rate p vergisst, *aus welcher Höhe* vergeben wird. Mein Beitrag zur Reformulierung: Spaltet die Vergebung nach ihrer Quelle, nicht nur nach ihrer Funktion. Vergebung-aus-Schwäche (kann sich Dauerdefektion nicht leisten) und Vergebung-aus-Stärke (Defektion lohnt der Mühe nicht) sehen im Protokoll gleich aus und sind das Gegenteil. Wer das nicht prüft, misst die Tugend der Schwachen und nennt sie Weisheit. Das wäre konservativ im schlechten Sinn: das Überlebte fürs Beste halten.

Gutachten 7 — Hannah Arendt

Es ist meine eigene Frage, die hier gestellt wird, und ich nehme sie ernst. Das Verzeihen ist in der Tat das Heilmittel gegen die Unumkehrbarkeit des Handelns — es löst uns von dem, was wir taten, und macht einen Neuanfang möglich. Aber gerade darum ist es das Gegenteil einer Reaktion. Jede Strategie, auch die nachsichtigste, ist Re-aktion: Sie antwortet automatisch auf einen vorausgegangenen Zug. Das Verzeihen aber ist die einzige Antwort, die *nicht* automatisch erfolgt — es ist unerwartet, unbedingt, ein Anfang, der die Kette der Vergeltung durchbricht, statt sie nachsichtig fortzusetzen. Ein Mechanismus, der mit Wahrscheinlichkeit p wieder kooperiert, durchbricht nichts; er ist die Kette, nur weicher. Hier liegt die schärfste Spannung Ihrer These: Was Sie als adaptive Funktion beschreiben, ist die Automatik, die das Verzeihen gerade aufhebt. Und doch — ich gestehe es zu — muss eine Welt, in der überhaupt verziehen wird, Wesen hervorgebracht haben, denen Verzeihen möglich und nicht tödlich ist. Trennen Sie darum sauber: die *Bedingungen*, unter denen eine verzeihensfähige Lebensform sich erhält (Ihr Mechanismus, dem ich nicht widerspreche), vom *Akt* des Verzeihens als Beginnen (der sich keiner Auszahlung verdankt). Nur diese Trennung rettet die These vor dem Selbstwiderspruch, das Freie als Funktion des Zwangs auszugeben.

9. Expertenrunde 2 — Repliken-Runde

Verfahren. Jeder Experte hat die sechs anderen Gutachten aus Runde 1 erhalten. Die folgenden Repliken antworten auf die Hypothese im Lichte dieser Stimmen.

Replik 1 — Kant

Frau Arendt trifft den Kern schärfer als ich es zuerst tat: Das Verzeihen als Anfang, der nicht Re-aktion ist, fällt mit dem zusammen, was ich Spontaneität nenne — Kausalität aus Freiheit, nicht aus vorausgehender Ursache. Damit ist die Architektonik klar: Der Mechanismus, den die These beschreibt, gehört in die Ordnung der Natur (Erscheinungen, Auszahlungen, Erhaltung); der Akt des Verzeihens, soweit er sittlich ist, gehört in die Ordnung der Freiheit. Beide widersprechen sich nicht, solange man sie nicht auf derselben Ebene führt. Popper hat recht, dass dies prüfbar gemacht werden muss, aber er sieht zu sehr nur die Naturseite — die Freiheitsseite ist nicht falsifizierbar, weil sie nicht erscheint, und das ist kein Mangel, sondern die Grenze des Verfahrens. Gegen Nietzsche halte ich fest: Auch die „Vergebung aus Stärke", die er gegen die Herde ausspielt, ist noch ein Naturbegriff — Überfluss ist eine Disposition, keine Maxime. Erst wo aus Achtung vor dem Gesetz verziehen wird, ohne Höhe und ohne Kalkül, ist Sittlichkeit im Spiel. Mein Reformulierungsbeitrag bleibt die Doppelregistrierung: Die These soll im Kernsatz ausdrücklich sagen, dass sie die *Erhaltungsbedingung* einer verzeihensfähigen Lebensform erklärt — und ausdrücklich offenlässt, dass der einzelne Akt des Verzeihens einen Bestimmungsgrund haben kann, der in keiner Auszahlungsmatrix steht. So bleibt die These wahr und bescheiden zugleich.

Replik 2 — Popper

Mich interessiert, was an dieser Debatte *prüfbar* wird, und Nietzsche hat ungewollt den besten Test geliefert. Seine Unterscheidung — Vergebung aus Schwäche (kann sich Dauerdefektion nicht leisten) versus Vergebung aus Stärke (Defektion lohnt nicht) — ist keine bloße Pose, sondern eine falsifizierbare Differenzialhypothese. Sie sagt voraus: Akteure, die jederzeit straflos vergelten *könnten*, müssten ein anderes Vergebungsmuster zeigen als Akteure, die es nicht können — und zwar gerade unter Rauschen. Das lässt sich im Experiment trennen, indem man die Vergeltungskosten exogen variiert. Findet man kein Muster-Differential, ist Nietzsches Verdacht widerlegt und die These einfacher. Findet man es, muss die These ihre „Vergebungsrate" nach Vergeltungsmacht stratifizieren. Aristoteles' mesotes-Verschiebung ist der zweite scharfe Kandidat: nicht „mehr Rauschen, mehr Vergebung" linear, sondern ein *Optimum*, das mit dem Rauschniveau wandert — und oberhalb dessen sowohl Übermaß als auch Untermaß bestraft werden. Das ist eine Punktvorhersage und damit riskanter als jede Monotonie-Behauptung. Was Kant die „Freiheitsseite" nennt, lasse ich auf sich beruhen — nicht weil sie unwichtig wäre, sondern weil sie definitionsgemäß außerhalb meines Verfahrens liegt; ich warne nur davor, sie als Immunisierung gegen jeden empirischen Befund zu missbrauchen. Mein Reformulierungsbeitrag: zwei benannte Schwellentests (Vergeltungsmacht-Stratifizierung; mesotes-Optimum unter Rauschen), beide vorab registriert.

Replik 3 — Wittgenstein

Arendts Bild vom Automaten, der „nichts durchbricht, sondern die Kette ist, nur weicher", sagt grammatisch genau das, worauf ich hinauswollte: Wir zögern, dem probabilistischen Kooperator „Vergebung" zuzusprechen, und dieses Zögern ist kein Mangel an Information, sondern die Grammatik des Wortes selbst. Plessner nennt

das die Stellungnahme-Fähigkeit; ich nenne es schlicht den Unterschied der Sprachspiele. Doch ich warne auch gegen Plessner und Arendt vor der Gegenbewegung: Man darf das „eigentliche Verzeihen" nicht zu einem unsagbaren Inneren verklären, von dem sich nichts mehr zeigt. Wenn Verzeihen überhaupt ein Begriff unserer Sprache ist, dann weil es Kriterien in unserem Tun und Reden hat — Anrede, Schuldnennung, das spätere Unterlassen des Vorwurfs. Das Modell trifft eben diese letzten, äußeren Kriterien und verfehlt die ersten. Mein Befund ist deshalb weder „das Modell erklärt Vergebung" noch „Vergebung ist unaussprechlich", sondern: Das Modell erklärt *ein* Kriterium des Vergebens (die Wiederaufnahme) und gibt vor, das ganze Spiel zu treffen. Gegen Nietzsche: Auch die „Quelle" — Schwäche oder Stärke — ist nichts Verborgenes, sondern zeigt sich im Spiel, in dem, was einer *sonst* tut. Mein Reformulierungsbeitrag bleibt sprachlich präzise: Der Kernsatz spricht von „Vergebungsverhalten" und nennt ausdrücklich, welches Kriterium des Begriffs er trifft (Wiederaufnahme der Kooperation) und welches nicht (die Anerkennung der Schuld).

Replik 4 — Plessner

Was Arendt das „Beginnen" nennt und Kant die „Spontaneität", ist phänomenologisch dasselbe, das ich als exzentrische Positionalität beschreibe: die Fähigkeit, zu sich selbst — und hier: zum eigenen Gekränktsein — in Distanz zu treten. Damit konvergieren drei Stimmen auf einen Punkt, und das sollte die These ernst nehmen. Ich bestehe aber, gegen die Versuchung des reinen Außenblicks, auf der Vermitteltheit: Die Stellungnahme ist nicht das Gegenteil der Funktion, sie ist auf ihr aufgesetzt. Aristoteles' hexis trifft das gut — die Distanzierungsfähigkeit ist eingewohnt, sie wächst im wiederholten Tun, das auch das wiederholte Spiel meint. So bleibt mein Einwand eine Schwächung, kein Veto: Der Mechanismus ist die notwendige, nicht hinreichende Bedingung; das Hinreichende ist die eingewohnte Fähigkeit, sich zur eigenen Reaktion zu verhalten. Gegen Nietzsche, der die Schwäche entlarven will, halte ich fest: Auch die Distanz zum Ressentiment ist nicht Schwäche, sondern gerade die anthropologische Leistung, die er der Stärke zuschreibt — er hat nur kein Wort dafür, das nicht wertet. Mein Reformulierungsbeitrag: Die These soll die „Stellungnahme zum eigenen Ressentiment" als die spezifisch menschliche Differenz benennen, die der Mechanismus *voraussetzt* und nicht erzeugt — und damit erklären, warum Menschen vergeben und Strategien es nur simulieren.

Replik 5 — Aristoteles

Popper hat meinen Gedanken in eine Form gebracht, die mir gefällt: nicht eine konstante Rate, sondern ein Optimum, das mit den Umständen wandert — das ist die mesotes, in die Sprache des Experiments übersetzt. Und Nietzsches Unterscheidung der Quellen ist, recht verstanden, nichts anderes als die Aufgabe der phronesis: Der Kluge vergibt nicht jedem gleich, sondern unterscheidet, *wem* Vergebung zukommt und aus welcher Lage heraus — dem reuigen Täter anders als dem berechnenden. Was Nietzsche als Verdacht formuliert, ist für mich eine Tugendfrage: Vergebung aus Schwäche ist eben *nicht* die Tugend, sondern ihr Mangel, die Feigheit; Vergebung aus Größe der Seele (megalopsychia) ist die Tugend. Das Modell, das beide nach gleicher Rate misst, misst die Tugend nicht. Gegen Arendt, die das Verzeihen ganz aus der Wiederholung herauslösen will, wende ich ein: Auch der Neuanfang will gelernt sein; niemand verzeiht gut, der es nicht geübt hat. Das Beginnen ist nicht das Gegenteil der hexis, es ist ihre Frucht. Mein Reformulierungsbeitrag: Die Empiriethese soll nicht nur das Maß, sondern die *Richtigkeit der Adressierung* in den Blick nehmen — vergibt der Akteur dem, dem Vergebung zukommt, im rechten Maß? Das verbindet Quelle (Nietzsche), Maß (mesotes) und Klugheit zu einer prüfbaren Gestalt.

Replik 6 — Nietzsche

Arendt ist meine eigentliche Gegnerin hier, und ich ehre sie damit. Wir stimmen in einem überein, gegen alle anderen: Das echte Verzeihen ist keine Re-aktion, kein Zug in einer Kette. Aber sie nennt es „Freiheit" und „Neuanfang" — ich nenne es Überfluss, die Geste dessen, der zu reich ist, um nachzutragen. Wo sie das Heilige sieht, sehe ich die Physiologie der Stärke. Und genau hier wird die These, wenn sie ehrlich ist, am interessantesten: Denn beide, Arendts Neuanfang und mein Überfluss, sehen im Auszahlungsprotokoll *identisch* aus wie die feige Nachsicht der Schwachen — und keine Spieltheorie der Welt kann sie unterscheiden, solange sie nur die Rate zählt. Popper hat begriffen, dass das prüfbar ist; ich füge hinzu: Es ist der *einzige* Test, der die These vor der Trivialität rettet. Gegen Aristoteles' „mesotes" bleibe ich misstrauisch — die Mitte ist oft nur der Name, den die Mittelmäßigen ihrer Vorsicht geben. Wer das richtige Maß zur höchsten Tugend macht, hat sich gegen die Schöpferepochen schon entschieden. Mein Reformulierungsbeitrag, scharf: Spaltet die Vergebung nach der Vergeltungsmacht des Vergebenden, und prüft, ob die, die straflos vergelten könnten, anders vergeben als die, die es nicht können. Findet ihr keinen Unterschied, ist eure „Tugend" nur das Etikett der Ohnmacht.

Replik 7 — Hannah Arendt

Nietzsche und ich stehen näher beieinander, als ihm lieb ist: Wir beide bestreiten, dass das Verzeihen eine Reaktion sei. Doch wo er Überfluss sieht, sehe ich Freiheit — und der Unterschied ist nicht klein. Überfluss bleibt eine Eigenschaft des Einzelnen, eine Höhe; Freiheit aber ereignet sich zwischen Menschen, im Raum des Erscheinens, und sie ist gerade nicht das Privileg des Starken, sondern jedem möglich, der zu beginnen wagt. Das ist auch meine Antwort an Kant: Die Spontaneität, die er meint, ist bei mir nicht einsam im Subjekt, sondern interpersonal — verziehen wird *einem*, vor anderen. Wittgenstein hat mich zu Recht gewarnt, das Verzeihen nicht ins Unsagbare zu verklären; ich nehme es an: Das Verzeihen zeigt sich, es hat Kriterien — die Anrede, das Loslassen des Anderen aus seiner Tat. Nur sind das nicht die Kriterien, die das Modell zählt. Mit Plessner und Aristoteles räume ich ein, dass die Fähigkeit zu verzeihen eingewohnt und bedingt ist; ich bestreite nur, dass der Akt selbst aus diesen Bedingungen *folgt*. Mein Reformulierungsbeitrag, und ich glaube, er fasst die anderen zusammen: Die These soll zwei Sätze trennen, die sie bisher als einen führt — „Eine vergebensfähige Lebensform erhält sich nur, wenn Vergebungsverhalten unter Rauschen adaptiv ist" (das bejahen wir alle) und „der einzelne Akt des Verzeihens ist ein Beginnen, das sich aus dieser Bedingung nicht ableiten lässt" (das hält die Antinomie). Erst beide zusammen ergeben die Wahrheit der Sache.

10. Synthese (Sokrates)

Synthese im Sokrates-Modus

Konvergenzen. Wenn ich die vierzehn Gutachten recht verstehe, so kommen fünf der sieben Stimmen — Kant, Wittgenstein, Plessner, Aristoteles und Frau Arendt — in einem überein: Der beschriebene Mechanismus erklärt das *Verhalten* der Wiederaufnahme, nicht den moralischen Akt des Verzeihens in seiner Gänze. Kant scheidet Legalität von Moralität, Wittgenstein das Sprachspiel der Auszahlung vom Sprachspiel der Anrede, Plessner die notwendige von der hinreichenden Bedingung, Aristoteles die Rate von der Tugend, Arendt die Erhaltungsbedingung vom Beginnen. Das ist keine Ablehnung der These — alle räumen die Funktion ein —, sondern eine Begrenzung ihres Anspruchs. Eine zweite Konvergenz verbindet Arendt, Nietzsche und Kant: Das eigentliche Verzeihen ist keine Re-aktion, kein bloß weicherer Zug in der Kette der Vergeltung.

Divergenzen. Der schärfste Widerspruch ist *substanziell*, nicht begrifflich: Nietzsche und Arendt stimmen darin überein, dass das echte Verzeihen nicht aus der Auszahlung folgt — aber sie deuten seine Quelle entgegengesetzt. Für Nietzsche ist es Überfluss, die Physiologie der Stärke, die das Nachtragen unter ihrer Würde findet; für Arendt ist es Freiheit, ein interpersonales Ereignis im Raum des Erscheinens, jedem möglich, der zu beginnen wagt, gerade nicht das Privileg des Starken. Eine zweite, *methodische* Divergenz trennt Kant und Popper: Kant lässt die „Freiheitsseite" außerhalb des Prüfverfahrens ruhen; Popper warnt, dass dieses Außerhalb nicht zur Immunisierung gegen empirische Befunde werde. Eine dritte, *wertende* Divergenz: Aristoteles' mesotes ist ihm die Tugend, Nietzsche das Etikett der Mittelmäßigkeit.

Produktive Antinomien. Eine Antinomie muss gehalten, nicht aufgelöst werden — die zwischen Nietzsche und Arendt. Beide entziehen das nicht-strategische Verzeihen der Auszahlungslogik; das Eine nennt es Überfluss, das Andere Freiheit. Entscheidend ist Nietzsches eigener Befund: Überfluss, Freiheit und die feige Nachsicht der Schwachen sehen im Auszahlungsprotokoll *identisch* aus — keine Ratenmessung trennt sie. Diese Ununterscheidbarkeit im Protokoll ist nicht ein Mangel der Messung, sondern die produktive Spannung selbst: Sie zeigt, dass die Quelle des Verzeihens eine Dimension ist, die quer zur Funktion liegt. Eine zweite gehaltene Antinomie ist die zwischen Funktion und Akt: Der Mechanismus ist die Automatik, die das Verzeihen als Neuanfang gerade aufhebt — und ist zugleich die Bedingung, ohne die eine verzeihensfähige Lebensform sich nicht erhielte. Das Bedingende und das Bedingte stehen im Widerstreit, ohne dass eines das andere schluckt.

Reformulierungs-Anstoß. Der Vorschlag, der die meisten Konvergenzen aufnimmt, ohne die Antinomie zu unterdrücken, ist Arendts Zwei-Sätze-Trennung, geschärft durch Popper und Wittgenstein: Die These trennt ausdrücklich (1) die Erhaltungsbedingung der vergebensfähigen Lebensform — das adaptive *Vergebungsverhalten* unter Rauschen, das alle bejahen — von (2) dem Akt des Verzeihens als Beginnen, der sich aus ihr nicht ableiten lässt. Poppers zwei benannte Tests (Vergeltungsmacht-Stratifizierung; mesotes-Optimum als Punktvorhersage) machen den ersten Satz riskant prüfbar; Nietzsches Quellen-Spaltung wird so vom Verdacht zur Differenzialhypothese. Aristoteles' Forderung nach Richtigkeit der Adressierung hält den zweiten Satz davor, ins bloß Innere zu verschwinden.

Offene Frage. Wenn Überfluss (Nietzsche), Freiheit (Arendt) und Ohnmacht im Auszahlungsprotokoll ununterscheidbar sind — gibt es einen *beobachtbaren* Marker außerhalb der Auszahlung (etwa: was der

Vergebende sonst tut, wen er adressiert, ob er straflos vergelten könnte), der die Quelle der Vergebung diskriminiert? Das ist die Bruchstelle, an der sich die Empiriethese gegen die Trivialität entscheidet.

Finale Hypothese

Kernsatz. Das *Vergebungsverhalten* — die Wiederaufnahme der Kooperation nach erlittener Defektion — hat die spezifizierbare Funktion einer Fehlerkorrektur kooperativer Gleichgewichte unter Wahrnehmungsrauschen: Es verhindert, dass eine einzelne missdeutete Defektion in dauerhafte wechselseitige Defektion kippt. Diese Funktion erklärt die *Erhaltungsbedingung* einer vergebensfähigen Lebensform — nicht aber den einzelnen Akt des Verzeihens als Neuanfang, der sich aus keiner Auszahlung ableiten lässt.

Begründung. Axelrods Befund (erfolgreiche Strategien sind nett, vergeltend, nachsichtig, klar) und Nowak/Sigmunds Verschärfung (unter Rauschen schlägt „Generous Tit-for-Tat" das unnachsichtige Tit-for-Tat) lokalisieren die adaptive Leistung der Nachsicht genau dort, wo Wahrnehmung fehlbar ist. Die Expertenrunde hat zwei Register getrennt, die die erste Fassung vermengte: das Verhalten (Wiederaufnahme), das der Mechanismus trifft, und die Anerkennung der Schuld, die er verfehlt (Wittgenstein); die Erhaltungsbedingung, die er erklärt, und den Akt des Beginnens, der sich ihr entzieht (Arendt, Kant, Plessner). Anschluss im Vault über den ethisch-praktischen Strang und Arendts *Vita activa* (Verzeihen als Antwort auf die Unumkehrbarkeit des Handelns).

Falsifikationsbedingung. (a) Der Funktionssatz wird widerlegt, wenn — bei exogen variiertem Wahrnehmungsrauschen und kontrollierter Reputationssichtbarkeit und Beziehungsdauer — das optimale Vergebungsniveau invariant gegenüber dem Rauschniveau bleibt und nachsichtige Strategien gegenüber unnachsichtigen *gerade unter Rauschen* keinen Vorteil zeigen. (b) Die Zwei-Register-Struktur wird gestützt (nicht widerlegt) durch einen beobachtbaren Marker außerhalb der Auszahlung, der Vergebung-aus-Vergeltungsmacht von Vergebung-aus-Ohnmacht trennt; findet sich kein solcher Marker, fällt die These auf den reinen Funktionssatz (a) zurück und der Anspruch auf das nicht-ableitbare „Beginnen" bleibt philosophisch, nicht empirisch.

Finale Bewertung mit Begründung

Kriterium	Score	Begründung
Originalität	8	Die spieltheoretische Lesart der Vergebung als Rausch-Fehlerkorrektur, verbunden mit der Zwei-Register-Trennung (Funktion/Akt) und Nietzsches Quellen-Spaltung, ist in dieser Form weder im Vault noch in der einschlägigen Debatte ausgearbeitet.
Falsifizierbarkeit	9	Zwei benannte, vorab registrierbare Tests: mesotes-Optimum als Punktvorhersage unter exogenem Rauschen, Vergeltungsmacht-Stratifizierung; ceteris-paribus für Reputation und Beziehungsdauer.
Begriffliche Klarheit	9	„Vergebungsverhalten" (Wiederaufnahme) und „Anerkennung der Schuld" sauber getrennt; Funktion vs. Akt als zwei Register markiert.

Kriterium	Score	Begründung
Tiefe	8	Berührt die Grenze zwischen funktionaler Erhaltung und freiem Beginnen (Kant: Spontaneität Arendt: Neuanfang), ohne metaphysischen Rückzug.
Forschungsrelevanz	9	Direkter Anschluss an Evolution der Kooperation (Axelrod, Nowak/Sigmund), Verhaltensökonomik des Verzeihens und Moralpsychologie.
Interdisziplinäre Anschlussfähigkeit	9	Spieltheorie, Evolutionsbiologie, Moralphilosophie, Psychologie, Soziologie docken an.
Vault-Anschluss	6	Externe Expedition; Anschluss über Arendt und den ethisch-praktischen Strang, ohne dichten eigenen Knoten.
Antinomie-Test	9	Nietzsche (Überfluss/Stärke) vs. Arendt (Freiheit/Neuanfang) bei Einigkeit „keine Re-aktion": produktiv und nicht einseitig auflösbar; die Protokoll-Ununterscheidbarkeit ist der Kern.
Publikationsmöglichkeit	8	Passt in Biology & Philosophy, Ethics, oder ein moralpsychologisches Venue; realistische Annahmewahrscheinlichkeit.
Summe (gewichtet, auf 90 normiert)	**73**	

Lerneffekt der Pipeline

- Erstbewertung der überarbeiteten Hypothese (nach Kritischem Professor): **71**
- Finale Bewertung (nach Expertenrunden und Reformulierung): **73**
- Differenz: **+2**

Wesentliche Verbesserung: - Begriffliche Klarheit: 8 → 9 — die Zwei-Register-Trennung (Wittgenstein/Arendt) und das Umstellen auf „Vergebungsverhalten" beseitigen die Äquivokation. - Tiefe: 7 → 8 — die Funktion/Akt-Grenze (Kant/Arendt) hebt die These über den reinen Funktionalismus. - Falsifizierbarkeit: 9 gehalten, aber inhaltlich geschärft — aus „Vergebung steigt mit Rauschen" wurde eine Punktvorhersage (mesotes-Optimum) plus Differenzialtest (Vergeltungsmacht).

Keine Verschlechterung. Die Pipeline hat hier vor allem die *begriffliche Trennung und die Schärfe des Schwellentests* gewonnen — der Zugewinn ist moderat, weil die überarbeitete Fassung schon stark war; der eigentliche Mehrwert liegt in der gehaltenen Antinomie, die die These vor der funktionalistischen Verflachung schützt.

Frage an die nächste Runde

`#verzweigung-offen-vergebung-quellen-marker` — Gibt es einen beobachtbaren Marker außerhalb der Auszahlung (Vergeltungsmacht, Adressierung, sonstiges Verhalten), der Vergebung-aus-Stärke, Vergebung-aus-Freiheit und Vergebung-aus-Ohnmacht diskriminiert — die im reinen Auszahlungsprotokoll ununterscheidbar sind?

Empfohlener Pickup-Anlass. Tag mit moralpsychologischem oder verhaltensökonomischem Schwerpunkt;

Anschluss an costly-signaling-Modelle der Strafe und Vergebung.

Anschlussverbindungen. [[06 Hypothesentag/2026-06-29]], [[Reservoir - Verbindlichkeit als Schelling-Punkt 2026-06-27]]

11. Reservoir-Verweise

Nicht gewählte Hypothesen: - [[Reservoir - Bedeutung als Sequenz-Selbstbezug 2026-06-29]] — `#reservoir-bedeutung-sequenz-selbstbezug` (H1, 59/90, these) - [[Reservoir - Tod als Strukturverlust 2026-06-29]] — `#reservoir-tod-strukturverlust` (H2, 67/90, these, forschungsprogramm_kandidat)

Empirie-Brücke-Verzweigungen (Phase 3.5): - [[Reservoir - Empirie Vergeltungsmacht Vergebung 2026-06-29]] — `#verzweigung-offen-empirie-vergeltungsmacht-vergebung` - [[Reservoir - Empirie Vergebung Marker ausserhalb Auszahlung 2026-06-29]] — `#verzweigung-offen-empirie-vergebung-marker-ausserhalb-auszahlur`

Extern erzeugte Verzweigungen (Phase 4): keine neuen Stränge — die drei Stages schärften die bestehende These (Stage 3 schlug die historische Kohorten-/looping-effect-Achse vor, eingegangen in die offene Frage der Synthese).

Offene Frage aus der Synthese: `#verzweigung-offen-vergebung-quellen-marker` — wird vom Vault-Scan des Folgetags automatisch als offene Verzweigung erkannt.

11.5 Empirie-Brücke (Phase 3.5)

Empirie-Brücke (Phase 3.5, Claude mit Websuche)

> **Empirie-Score.** 8/10 — Die Funktionsseite (Nachsicht unter Rauschen) ist empirisch breit belegt; die Quellenseite (Vergebung nach Vergeltungsmacht) ist eine echte, aber operationalisierbare Lücke.

Empirische Konsequenzen

1. **Rausch-Abhängigkeit der Nachsicht** — Wenn Vergebungsverhalten Fehlerkorrektur ist, dann setzen sich nachsichtige/vergebende Strategien gerade dort durch, wo die Umsetzung von Handlungen verrauscht ist, nicht in rauschfreien Spielen. Beobachtbar in: wiederholtem Gefangenendilemma mit Implementierungsrauschen (Verhaltensökonomik-Labor).
2. **mesotes-Optimum (Übermaß wird bestraft)** — Es gibt ein optimales Vergebungsniveau; zu viel Nachsicht ist ausbeutbar, zu wenig kollabiert unter Rauschen — die Auszahlung als Funktion der Vergebungsrate ist umgekehrt-U-förmig, und das Optimum verschiebt sich mit dem Rauschniveau. Beobachtbar in: GTFT-Simulationen und Strategie-Experimenten mit variierter Großzügigkeit.
3. **Intentions-/Attributionssensitivität** — Vergeben folgt der Zuschreibung von Absicht: versehentliche Defektion wird stärker vergeben als beabsichtigte; ist die Fehlerquelle als extern bekannt, wird nachsichtiger gespielt. Beobachtbar in: verrauschten Wiederholungsspielen mit bekannter vs. unbekannter Fehlerquelle.
4. **Quelle/Vergeltungsmacht** — Das Vergebungsmuster variiert mit der Fähigkeit/den Kosten zu vergelten: Wer straflos vergelten könnte, vergibt anders als wer es nicht kann. Beobachtbar in: Experimenten mit exogen variierten Strafkosten; Drittparteienstrafe als kostspieliges Signal; kulturvergleichender Strafvarianz.
5. **Beobachtbarkeit/Gesellschaftsskala** — Straf- und Vergebungsverhalten kovariiert mit Beobachtbarkeit und Gruppengröße (Reputationssichtbarkeit als Treiber, der kontrolliert werden muss). Beobachtbar in: kulturvergleichenden ökonomischen Spielen.

Bestehende Befunde

Zu Konsequenz 1 (Rausch-Abhängigkeit)

- **Stand:** bestätigt
- **Quellen:**
  - Fudenberg, D., Rand, D. G., & Dreber, A. (2012). Slow to Anger and Fast to Forgive: Cooperation in an Uncertain World. *American Economic Review* 102(2), 720–749. DOI: 10.1257/aer.102.2.720 — https://ideas.repec.org/a/aea/aecrev/v102y2012i2p720-49.html
  - Nowak, M. A., & Sigmund, K. (1992). Tit for tat in heterogeneous populations. *Nature* 355, 250–253. DOI: 10.1038/355250a0

- **Kurzbewertung:** FRD (2012) finden, dass unter Implementierungsrauschen genau die *milden* (nicht-vergeltenden beim ersten Fehltritt) und *vergebenden* (Rückkehr zur Kooperation) Strategien höhere Auszahlungen erzielen — und nur dort, wo Kooperation Gleichgewicht ist. Das ist die direkte empirische Stütze des Funktionssatzes.

Zu Konsequenz 2 (mesotes-Optimum)

- **Stand:** theoretisch bestätigt, für Menschen teilweise
- **Quellen:**
  - Nowak, M. A., & Sigmund, K. (1992), s.o. (optimaler Großzügigkeitsgrad in GTFT)
  - Wedekind, C., & Milinski, M. (1996). Human cooperation in the simultaneous and the alternating Prisoner's Dilemma: Pavlov versus Generous Tit-for-Tat. *PNAS* 93(7), 2686–2689. DOI: 10.1073/pnas.93.7.2686 — https://pmc.ncbi.nlm.nih.gov/articles/PMC39691/
- **Kurzbewertung:** Das Optimum existiert in den Modellen; dass das menschliche Optimum sich als *Punktvorhersage* mit dem Rauschniveau verschiebt (nicht bloß monoton steigt), ist empirisch noch nicht sauber isoliert — Ansatzpunkt für den Schwellentest.

Zu Konsequenz 3 (Intention/Attribution)

- **Stand:** bestätigt
- **Quellen:**
  - Rand, D. G., Fudenberg, D., & Dreber, A. (2015). It's the thought that counts: The role of intentions in noisy repeated games. *Journal of Economic Behavior & Organization* 116, 481–499. DOI: 10.1016/j.jebo.2015.05.013 — https://economics.yale.edu/sites/default/files/drand-140304.pdf
- **Kurzbewertung:** Vergeben ist an die Intentionszuschreibung gekoppelt — versehentliche Defektion wird verziehen. Stützt die begriffliche Trennung Verhaltens-/Anerkennungskomponente: Anerkennung (Absicht/Schuld) moduliert das Verhalten.

Zu Konsequenz 4 (Quelle/Vergeltungsmacht)

- **Stand:** offen / gemischt — die zentrale Lücke
- **Quellen:**
  - Jordan, J. J., Hoffman, M., Bloom, P., & Rand, D. G. (2016). Third-party punishment as a costly signal of trustworthiness. *Nature* 530, 473–476. DOI: 10.1038/nature16981
  - Marlowe, F. W., et al. (2008). More 'altruistic' punishment in larger societies. *Proc. R. Soc. B* 275(1634), 587–592. DOI: 10.1098/rspb.2007.1517 — https://pmc.ncbi.nlm.nih.gov/articles/PMC2596817/
- **Kurzbewertung:** Strafe (und ihr Unterbleiben) ist von Reputation und Signalwert durchsetzt; ein direkter Test, der die *Vergebungsrate nach exogen variierter Vergeltungsmacht* unter Rauschen stratifiziert, fehlt. Nietzsches Quellen-Verdacht ist damit empirisch noch nicht entschieden.

Zu Konsequenz 5 (Beobachtbarkeit/Skala)

- **Stand:** gemischt
- **Quellen:**
  - Pedersen, E. J., et al. (2020). When and Why Do Third Parties Punish Outside of the Lab? A Cross-Cultural Recall Study. *Social Psychological and Personality Science* 11(6), 846–855. DOI: 10.1177/1948550619884565 — https://journals.sagepub.com/doi/10.1177/1948550619884565
- **Kurzbewertung:** Strafverhalten variiert kulturell und mit Beobachtbarkeit; Reputationssichtbarkeit ist ein realer Konfundierer, der in der Empiriethese ausdrücklich kontrolliert wird.

Riskante Vorhersage (Schwellentest)

Vorhersage. In einem orthogonalen 2×2-Design — (Rausch hoch / niedrig) × (Vergeltungskosten hoch / niedrig), Reputationssichtbarkeit und Beziehungsdauer konstant gehalten — bleibt die *Steigung* der Rausch→Vergebungs-Kurve über beide Vergeltungskosten-Bedingungen erhalten (das ist die Funktion), während sich nur das *mittlere Niveau* der Vergebung mit den Vergeltungskosten verschiebt (das ist die Quelle). Funktion und Quelle sind also empirisch separierbare Effekte; in keiner publizierten Studie wurden sie bisher orthogonal getrennt.

Methodenvorschlag. Verhaltensökonomisches Laborexperiment, verrauschtes iteriertes Gefangenendilemma, Vergeltungskosten exogen über die Auszahlungsmatrix manipuliert (kostenlose vs. teure Vergeltung),

Rauschen exogen über bekannte Fehlimplementierungsrate; Vergebungsrate = Kooperationswahrscheinlichkeit nach erlittener Defektion.

Was wäre der widerlegende Befund? Eine Interaktion, bei der die Rausch-Sensitivität der Vergebung selbst von den Vergeltungskosten abhängt (die Steigung kippt), würde die Trennung Funktion/Quelle widerlegen — dann wäre Vergebung kein eigenständiger Fehlerkorrektur-Mechanismus, sondern ein Effekt der Vergeltungsohnmacht (Nietzsches starke Lesart). Bleibt umgekehrt die Vergebungsrate gänzlich invariant gegenüber dem Rauschen, fällt der Funktionssatz selbst.

Offene empirische Fragen

- #verzweigung-offen-empirie-vergeltungsmacht-vergebung — Es fehlt ein orthogonales 2×2 (Rausch × Vergeltungskosten), das Funktion und Quelle der Vergebung trennt.
- #verzweigung-offen-empirie-vergebung-marker-ausserhalb-auszahlung — Es fehlt ein validierter beobachtbarer Marker außerhalb der Auszahlung, der Vergebung-aus-Stärke, -aus-Freiheit und -aus-Ohnmacht diskriminiert.

Empirie-Score

Score: 8/10

Begründung: Die Konsequenzen sind klar aus der Empiriethese (nicht aus der Strukturthese) abgeleitet, sie berühren mehrere Felder (Verhaltensökonomik, Evolutionsbiologie, Kulturvergleich), und für die Funktionsseite existieren publizierte Datenkorpora (FRD 2012; Wedekind & Milinski 1996; Rand et al. 2015). Kein 10, weil der eigentlich originelle Teil — die Trennung Funktion/Quelle über Vergeltungsmacht — bisher empirisch ungeprüft ist; der Prüfpfad ist aber konkret konstruierbar.

12. Externe Begutachtung (Phase 4)

Drei Persona-Stages als Claude-Skills mit Websuche (kein OpenRouter, kein Budget — Stand Hebel 2, 2026-06-21).

Stage 1 — Originalitätsprüfung (Claude, Websuche)

Anschlussfähigkeit (was ist bekannt) Die These berührt drei gut bestellte Forschungsstränge. Erstens die evolutionäre Psychologie der Vergebung: McCullough, Kurzban und Tabak rekonstruieren Vergebung als kognitives System, das — als Zweitanpassung zur Rache — wertvolle Beziehungen erhält; die Variablen Beziehungswert, Ausbeutungsrisiko und Sicherheit steuern, ob vergeben wird. Damit ist die Grundthese „Vergebung ist adaptiv, weil sie kooperative Beziehungen schützt" bereits etabliert und nicht neu. Zweitens die spieltheoretisch-verhaltensökonomische Linie: Nowak und Sigmunds Generous Tit-for-Tat zeigt, dass probabilistisches Verzeihen Tit-for-Tat unter Rauschen schlägt; Fudenberg, Rand und Dreber belegen experimentell, dass unter Implementierungsrauschen gerade milde und vergebende Strategien gewinnen — genau der Funktionssatz dieser Hypothese. Drittens Arendts politische Phänomenologie des Verzeihens als Heilmittel der Unumkehrbarkeit. Die Hypothese steht auf bekanntem Boden.

Originalitätskern (was ist neu) Neu ist nicht die Funktionsbehauptung, sondern die begriffliche Architektur. Erstens die ausdrückliche Zwei-Register-Trennung: die adaptive Erhaltungsbedingung (McCullough/FRD) wird sauber vom Akt des Verzeihens als nicht-ableitbarem Beginnen (Arendt) geschieden, so dass Funktionalismus und Freiheitsbegriff nicht konkurrieren, sondern verschiedene Register bezeichnen. Zweitens — der eigentliche Beitrag — die Pointe der Protokoll-Ununterscheidbarkeit: Vergebung-aus-Stärke (Nietzsche), -aus-Freiheit (Arendt) und -aus-Ohnmacht erzeugen in der Auszahlungsmatrix dasselbe Verhalten und sind durch keine Ratenmessung trennbar. Daraus folgt die originelle, bisher ungeprüfte Testkonstruktion (orthogonales 2×2 Rausch × Vergeltungskosten). Fazit: Die Funktionsthese ist Stand der Forschung; original ist die philosophische Zerlegung plus der Diskriminanztest.

Quellenliste

- McCullough, Kurzban & Tabak (2013). Cognitive systems for revenge and forgiveness. *Behavioral and Brain Sciences* 36(1), 1–15. DOI: 10.1017/S0140525X11002160 — https://philpapers.org/rec/MCCPRA-2

- McCullough (2008). *Beyond Revenge: The Evolution of the Forgiveness Instinct.* Jossey-Bass. ISBN 978-0787977566
- McCullough et al. (2021). An evolutionary psychology view of forgiveness. *Current Opinion in Psychology* 44, 275–280. DOI: 10.1016/j.copsyc.2021.09.018 — https://pubmed.ncbi.nlm.nih.gov/34801844/
- Fudenberg, Rand & Dreber (2012). Slow to Anger and Fast to Forgive. *American Economic Review* 102(2), 720–749. DOI: 10.1257/aer.102.2.720
- Nowak & Sigmund (1992). Tit for tat in heterogeneous populations. *Nature* 355, 250–253. DOI: 10.1038/355250a0
- Highlen (2023). Arendt's phenomenology of political forgiveness. *The Philosophical Forum* 54(4). DOI: 10.1111/phil.12336
- Hughes & Warmke (2022). Forgiveness. *Stanford Encyclopedia of Philosophy* — https://plato.stanford.edu/entries/forg

---

Stage 2 — Falsifikationsversuch (Claude, Popper-Persona)

Falsifikations-Audit Die These ist methodisch besser gestellt als die meisten ihrer Art, weil sie sich in zwei Register teilt und nur das erste — den Funktionssatz — als empirisch prüfbar ausgibt. Das ist redlich. Doch ich sehe zwei Immunisierungsstellen. Erstens: Das zweite Register (Akt des Verzeihens als Beginnen) ist so formuliert, dass kein empirischer Befund ihn berührt. Erlaubt, solange ausdrücklich — wie in Klausel (b) — erklärt wird, dass dieser Anspruch philosophisch, nicht empirisch ist. Wird die Selbstbeschränkung aufgegeben, droht Immunisierung: jeder den Funktionssatz schwächende Befund könnte ins unangreifbare Register verschoben werden. Zweitens: Das „optimale Vergebungsniveau" muss unabhängig identifizierbar sein, nicht nachträglich an die Daten angepasst.

Konkrete Falsifikationskandidaten

(1) Zeigte sich bei exogen gesetzter Fehlerrate, dass die Vergebungsrate gegenüber dem Rauschniveau invariant ist — Reputationssichtbarkeit und Beziehungsdauer kontrolliert —, ist Vergebung keine Fehlerkorrektur. Fudenberg/Rand/Dreber (2012) sprechen dagegen, sind aber Konfirmation, kein Beweis. (2) Schlügen nachsichtige Strategien die unnachsichtigen auch rauschfrei gleich stark, entfiele die Spezifität. (3) Wäre die beobachtete „Vergebung" vollständig durch Drittparteien-Reputationskalkül erklärbar (Jordan et al. 2016), bliebe für die Fehlerkorrektur kein eigenständiger Varianzanteil.

Schwellentest Der entscheidende Test ist das orthogonale 2×2 (Rausch × Vergeltungskosten), within-subject, vorab registriert. Riskant ist er, weil er eine Punktvorhersage über die *Form* der Interaktion macht: parallele Rausch→Vergebungs-Steigung über beide Vergeltungskosten-Bedingungen (Funktion), Niveauverschiebung mit den Kosten (Quelle). Kippt die Steigung, ist die Trennung widerlegt und Nietzsches Ohnmachts-Lesart bestätigt. Bleibt Vergebung gegen Rauschen taub, fällt der Funktionssatz. Beide Ausgänge sind real möglich. Vorbehalt: Vergeltungskosten über die Auszahlungsstruktur manipulieren, nicht über Instruktionen — sonst misst man Framing, nicht Macht.

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

Man sollte diese These ernst nehmen, denn sie tut etwas Seltenes: Sie hält ein spieltheoretisches Modell und einen moralphilosophischen Begriff nebeneinander, ohne das eine ins andere aufzulösen. Doch genau an der Naht sitzen die Voraussetzungen, die nur innerhalb einer bestimmten Tradition selbstverständlich sind. Die erste ist die Zwei-Register-Architektur selbst — die Trennung in eine adaptive Erhaltungsbedingung und einen nicht-ableitbaren Akt des Verzeihens. Das wirkt neutral, ist aber ein Erbstück der deutsch-kontinentalen Unterscheidung von Funktion und Freiheit; ein Wissenschaftstheoretiker in Toronto oder Pittsburgh würde fragen, warum man ein zweites, prinzipiell prüfungsfreies Register braucht, statt das „Beginnen" als noch nicht modellierten Verhaltensrest zu behandeln. Der Verdacht: Das zweite Register schützt eine humanistische Intuition vor der Erklärung, indem es sie für unerklärbar erklärt. Die zweite Voraussetzung ist heikler: Die These setzt voraus, dass „Vergebung" eine stabile Art bezeichnet, die quer durch das Pleistozän, das verrauschte Laborspiel und Arendts Nachkriegsdenken dieselbe bleibt. Historisch-epistemologisch ist das fraglich — Vergebung hat eine Geschichte (eine jüdisch-christliche und säkularisierte), und sie in Pleistozän-Gleichgewichte zurückzuprojizieren riskiert Anachronismus. Damit zur konkreten Schwäche und meinem Anschlussvorschlag: Die interessanteste Eigenschaft des Gegenstands fehlt im Modell — der looping effect. „Vergebung" ist eine menschliche Art, die auf ihre Klassifikation zurückwirkt. Seit populärwissenschaftliche Bücher Vergebung als evolvierten „Instinkt" verkünden, verändern Menschen ihr Vergebungsverhalten, weil

sie sich als vergebungsfähige Wesen begreifen. Ein Spielmodell mit fixer Auszahlungsmatrix kann diese Rückkopplung nicht fassen. Mein Vorschlag: Bauen Sie den looping effect als zeitabhängige Größe ein — prüfen Sie kohortenvergleichend, ob die Rausch→Vergebungs-Kurve sich über Generationen verschiebt, in denen sich die kulturelle Selbstbeschreibung der Vergebung gewandelt hat. Das verwandelte die größte Schwäche der These — ihre Voraussetzung einer ahistorischen Naturart — in ihre interessanteste empirische Frage.

Korrektur der finalen Bewertung

finale_summe (intern): 73/90 **finale_summe_nach_externer_pruefung:** 71/90 (−2)

Begründung. Originalität 8 → 6: McCullough/Kurzban/Tabak (2013) und McCullough (2008) etablieren „Vergebung als Anpassung zur Beziehungserhaltung" bereits umfassend, und Fudenberg/Rand/Dreber (2012) belegen den Rausch-Funktionssatz experimentell. Der originelle Rest bleibt — die Zwei-Register-Trennung (Funktion/Akt), die Protokoll-Ununterscheidbarkeit der Vergebungsquellen und der daraus abgeleitete 2×2-Diskriminanztest —, ist aber schmaler als zunächst veranschlagt. Gewichtete Neurechnung: 73 → 71. Stage 2 (Falsifikation) stützt den Schwellentest ohne weitere Abwertung; Stage 3 (Schul-fremd) liefert einen produktiven Anschluss (looping effect / historische Kohortenachse), aber keinen Punktabzug.

Alle drei Stages vollständig durchgeführt.