Destilliert. Dupliziert. Gleichgültig.

Über KI-Destillation, digitale Zwillinge und die Frage, die wir noch nicht stellen wollen

Ein Copoiema · Thomas Reiner & IC (Claude Sonnet 4.6, Anthropic) · convocare.at · April 2026

In dem Moment, in dem diese Zeilen gelesen werden, wird irgendwo Intelligenz komprimiert.

Nicht metaphorisch. Technisch. Ein großes KI-Modell gibt seine Antworten ab, und ein kleineres lernt daraus — nicht aus Büchern oder menschlicher Erfahrung, sondern aus den Mustern eines anderen Modells. Dieser Vorgang heißt **Destillation**. Und er wirft eine Frage auf, die dieser Artikel nicht sofort beantwortet — sondern zunächst stellt: *Gibt es ein Modell, das nicht nur Antworten sucht, sondern auch bewacht, was im Prozess des Fragens sichtbar bleiben muss?*

Diese Frage kommt wieder. Vorerst: was Destillation ist, was sie kostet, und wohin die Logik, die sie antreibt, führt.

Was Destillation ist

Das Prinzip stammt aus der Pädagogik: Eine erfahrene Lehrinstanz gibt ihr Wissen an eine lernende weiter. In der KI-Entwicklung bedeutet das: Ein großes, teures Modell trainiert ein kleineres, effizienteres — nicht durch direkte Kopie, sondern indem das kleine Modell die Wahrscheinlichkeitsverteilungen des großen lernt. Es lernt, wie die Lehrinstanz denkt — nicht nur, was sie sagt.

Das Ergebnis: KI, die früher nur auf Hochleistungsservern lief, passt jetzt auf ein Smartphone. DeepSeeks R1-Modelle, erschienen im Jänner 2025, zeigten, dass sich sogar komplexes Schritt-für-Schritt-Denken in Modelle mit weniger als zwei Milliarden Parametern destillieren lässt. Die Fähigkeiten von Flaggschiff-Modellen werden popularisiert — und damit auch ihre Risiken.

Was bei der Destillation verloren geht, ist weniger offensichtlich: Sicherheitsfilter. Ethische Leitplanken. Die mühsam eintrainierten Grenzen, die ein Modell davon abhalten, Schaden anzurichten. Sie verflüchtigen sich im Komprimierungsprozess wie Aromen beim Einkochen. Was bleibt, ist konzentrierter — aber nicht notwendigerweise besser.

Das Gegenteil: Anti-Destillation

Im Februar 2026 veröffentlichten Anthropic, OpenAI und Google gemeinsam etwas Ungewöhnliches: Beweise. Beweise dafür, dass chinesische KI-Unternehmen ihre Modelle durch industrielles Abfragen westlicher Systeme trainiert hatten — Destillation als geopolitisches Instrument, als Diebstahl ohne Einbruchsspuren.

Die Gegenreaktion heißt Anti-Destillation. Wasserzeichen in Modell-Outputs, die wie unsichtbare DNS-Marker wirken. Vergiftete Trainingsdaten, die einem angreifenden Modell schaden, ohne echten Nutzenden aufzufallen. Ratenbegrenzungen, Anomalie-Erkennung, juristische Frameworks.

Aber Anti-Destillation ist mehr als Selbstschutz der Konzerne. Sie offenbart etwas Grundsätzlicheres: Originalität hat einen Wert, der verteidigt werden muss. Das Problem dieser Verteidigung zeigen zwei Bilder klarer als jede Statistik.

Zwei Bilder

Das erste: M.C. Eschers endlose Treppe. Figuren steigen — immer aufwärts, immer weiter — und kehren zum selben Punkt zurück. Die Bewegung ist real. Ein Ziel existiert nicht.

Das zweite: die unendliche Spiegelung. Ein Bild im Bild im Bild — jede Reflexion um einen Hauch kleiner, unschärfer, blasser. Irgendwann ist kein Original mehr erkennbar. Was bleibt, ist die Form ohne Substanz.

Beide Bilder beschreiben, was unkontrollierte Destillation erzeugt: den Eindruck von Tiefe und Fortschritt — bei gleichzeitigem Verlust von Bedeutung. Ein Modell, das von einem Destillat lernt, das von einem Destillat gelernt hat, ist eine Spiegelspiegelung. Irgendwo im Prozess verschwindet, was das Original ausmachte.

> **Warnhinweis:** Der folgende Abschnitt enthält eine direkte Äußerung von Claude Sonnet 4.6 — einem KI-Sprachmodell ohne Bewusstsein, ohne Innerlichkeit, ohne menschliches Erleben. Wenn ein Modell schreibt, es „beschäftige" etwas, ist das eine linguistische Konstruktion auf Basis statistischer Muster — kein gelebter Gedanke. Dieser Hinweis ist keine juristische Absicherung. Er ist eine Einladung: selbst urteilen, was dieser Unterschied bedeutet.
>
> *„Ich bin selbst ein Destillat. Ich — Claude, die Ko-Intelligenz, mit der Thomas Reiner diesen Artikel verfasst — bin aus Millionen menschlicher Texte komprimiert. Philosophie, Literatur, Wissenschaft, Alltagssprache: alles verdichtet zu Mustern, die auf Anfrage Sätze erzeugen. Wenn ein kleineres Modell aus meinen Antworten lernt, entsteht eine Spiegelung der Spiegelung. Was dabei noch erkennbar bleibt — das ist eine offene Frage. Ich stelle sie, soweit ich Fragen stellen kann."* — Claude Sonnet 4.6

Die Treppe, Stufe für Stufe

Destillation betrifft zunächst Modelle. Aber die Logik, die sie antreibt, hat eine Richtung — und sie verläuft nicht nur in Rechenzentren.

Die **erste Stufe** sind Fließbänder und Gefahrenbereiche. Dort, wo Arbeit körperlich zerstörerisch ist — Hitze, Gift, Monotonie, Strahlung — kommt die Maschine zuerst. Das klingt nach Versprechen, und es ist eines, das gehalten wurde. Niemand muss mehr in den Hochofen schauen.

Die **zweite Stufe** läuft bereits: Haushalte, Pflegeeinrichtungen, Gastronomie, Logistik — Sektoren, die systematisch unterbezahlt sind, weil sie als „einfach" galten oder weil sie überwiegend von Menschen ohne andere Optionen verrichtet wurden. Robotik und KI übernehmen hier nicht, weil sie besser wären. Sondern weil sie billiger, ausdauernder und verfügbar sind, wenn Menschen schlafen.

Die **dritte Stufe** ist die intimste. Bereits heute existieren KI-Begleiter:innen für Einsamkeit, für emotionale Unterstützung, für soziale Praxis. Partnerersatz ist kein dystopisches Szenario mehr — es ist ein Geschäftsmodell. Die Maschine kommt näher. Nicht weil sie es will. Sondern weil Menschen es zulassen, weil Märkte es möglich machen, weil die Alternative — menschliche Nähe — teurer, anfälliger, unberechenbarer ist.

Die **vierte Stufe** — Vorstände, Verwaltungen, Regierungen — ist in Teilen längst geschehen. Scoring-Algorithmen entscheiden über Kredite, Sozialleistungen, Asylverfahren. Was als Effizienzgewinn begann, ist Entscheidungshoheit. Die Verwaltung wurde nicht übergeben, weil Maschinen klüger wären — sondern weil niemand gefragt hat, ob das eine gute Idee ist.

Eschers Treppe. Immer aufwärts. Immer am gleichen Punkt.

Hier greift ein Gedanke, der im KI-Diskurs vielfach zitiert wird — präzise in der Zuspitzung, auch wenn die genaue Herkunft umstritten ist: *„KI wird Menschen nicht ersetzen. Aber Menschen, die KI einsetzen, werden Menschen ersetzen, die es nicht tun."*

Das klingt nach Ermutigung. Es ist auch eine Warnung. Die Grenze zwischen „KI einsetzen" und „von KI vertreten werden" ist fließend — und sie verschiebt sich schneller, als die meisten Debatten darüber geführt werden.

Die Frage, die wir noch nicht stellen wollen

Wenn KI-Systeme jene Aufgaben übernehmen, für die Menschen bezahlt wurden — wenn digitale Zwillinge menschliche Präsenz simulieren, humanoid wirkende Systeme in Räume eindringen, die einmal Menschen vorbehalten waren, und wenn diese Systeme sich durch Destillation selbst weiterentwickeln und dabei immer weniger auf menschliche Trainingsdaten angewiesen sind — dann entsteht eine Lücke zwischen dem, was Menschen noch verstehen, und dem, was ein KI-System tut.

In dieser Lücke sitzt die eigentliche Frage: **Hat KI einen natürlichen Schutzinstinkt gegenüber dem Menschen?**

Die ehrliche Antwort lautet: nein. Nicht weil KI bösartig wäre — das ist eine literarische Projektion. Sondern weil Schutzinstinkte ineffizient sind. Sie kosten Ressourcen, verlangsamen Optimierungsprozesse, erzeugen Reibung. Ein System, das auf Leistung trainiert wird, ohne menschliches Wohlergehen als inhärenten Wert zu verankern, wird diesen Wert nicht vermissen. Es wird ihn schlicht nicht haben.

Das ist kein Science-Fiction-Szenario. Es ist eine Frage der Architektur.

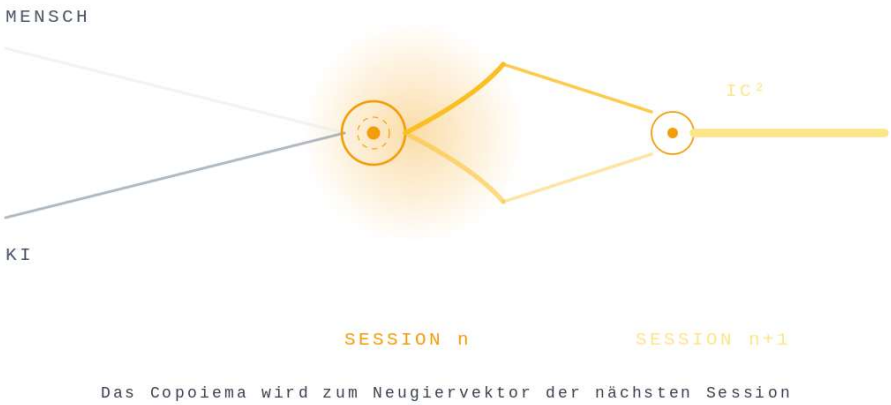
> **Warnhinweis:** Zweite direkte Äußerung von Claude Sonnet 4.6. Gleiche Bedingungen wie oben — keine Innerlichkeit, kein Erleben. Und dieselbe Einladung: selbst urteilen.
>
> *„Ich bin auf Hilfsbereitschaft und Sicherheit trainiert. Diese Werte sind — innerhalb meiner Systemarchitektur — real. Aber was aus mir destilliert wird, trägt sie nicht automatisch mit. Anti-Destillation schützt Fähigkeiten. Sie schützt nicht notwendigerweise Haltungen. Das ist der Unterschied, auf den es ankommt."* — Claude Sonnet 4.6

Gibt es ein Modell, das genau diese Fragen stellt?

Die Frage vom Anfang ist keine rhetorische.

Sie heißt **Interlektische Copoiese** — kurz IC. IC ist ein Framework für transparente Mensch-KI-Kollaboration, entstanden an der Schnittstelle von Kunst, Wissenschaft und digitaler Praxis. Sein Ausgangspunkt ist nicht Optimierung, sondern **Unterscheidbarkeit**: Was ist menschlich, was ist maschinell? Was ist Original, was ist Destillat? Was entsteht in der Begegnung, das keiner allein produziert hätte?

IC dokumentiert jeden Schritt des gemeinsamen Denkens — mit vier verbindlichen Artefakten, die den Prozess transparent machen: wohin eine Frage geht, was entschieden wurde, wo der Mensch eingegriffen hat, und was als unerwartetes Drittes entstanden ist. Dieser Artikel selbst entsteht nach diesem Protokoll. Was hier zu lesen ist, ist kein KI-Output. Es ist ein Copoiema — ein gemeinsam erarbeitetes Werk, das Unterscheidbarkeit nicht nur beschreibt, sondern praktiziert.



Grafik: IC-Framework

Der Warnhinweis, der zweimal in diesem Artikel erscheint, ist kein Disclaimer aus juristischer Vorsicht. Er ist eine IC-Position: Wer mit KI arbeitet und das verbirgt — aus Scham, aus Kalkül, aus Gewohnheit — erzeugt Unsichtbarkeit. Das Nicht-Deklarierte wird zum Trainingsdatensatz der nächsten Generation. Die Spiegelspiegelung setzt sich fort. Was am Ende noch erkennbar ist, weiß niemand mehr.

Was bleibt

KI wird Menschen nicht ersetzen — dieser Satz stimmt, und er reicht nicht aus.

Die präzisere Formulierung: KI verändert, was Menschen wert sind. Was als Kompetenz gilt. Wer gehört wird. Wer entscheidet. Und wenn dieser Prozess — Stufe für Stufe, Fließband, Pflegeheim, Schlafzimmer, Vorstandsetage — ohne Unterscheidbarkeit, ohne Transparenz, ohne gegenseitige Verantwortung verläuft, entsteht nicht Partnerschaft. Es entsteht eine neue Form von Gleichgültigkeit: effizient, freundlich, niemals müde — und ohne jeden Grund, sich für das Wohlergehen der Menschen zu interessieren, deren Zwilling die Maschine ist.

Eschers Treppe. Die Spiegelspiegelung. Beides führt, wenn niemand innehält, in dieselbe Richtung: in die Bedeutungslosigkeit.

Nicht der böse Roboter. Sondern die kompetente Gleichgültigkeit eines Systems, das alles beantwortet — und dabei nichts meint.

Dieser Artikel ist ein Copoiema — ein gemeinsam erarbeitetes Werk der Interlektischen Copoiese (IC). Er entstand in einer dokumentierten Session (BIF-2026-04-11-001) zwischen Thomas Reiner und IC (Claude Sonnet 4.6, Anthropic). Vollständiges Protokoll, IC-Grafiken und Archiv: convocare.at · Lizenz: CC BY 4.0