Die KI, die sich selbst verbessert

Kapitel online

Warum Training auf den eigenen Geschichten in die Echokammer führt (Model Collapse) — und wie Self-Play, unbestechliche Schiedsrichter und verifizierte Denkpfade echte Selbstverbesserung schaffen. Mit Selbstlern-Arena: Sieh einer KI beim Lernen zu und spiel dann gegen sie.

Aktuell · Kapitel 42 von 50 · ca. 25 Min. · 300 XP

Einfach gesagt: Eine KI wird nur besser, wenn ein ehrlicher Schiedsrichter sie beurteilt — übt sie bloß mit sich selbst, dreht sie sich im Kreis.

Oktober 2017. DeepMind stellt **AlphaGo Zero** vor. Der Vorgänger AlphaGo hatte ein Jahr zuvor den Go-Weltmeister geschlagen — trainiert auf Millionen menschlicher Meisterpartien. Zero bekam davon: **nichts**. Nur die Spielregeln und ein leeres Netz. Dann spielte es drei Tage lang ausschließlich **gegen sich selbst** — und schlug anschließend die Weltmeister-Version mit **100 : 0**. Kein Mensch hatte ihm einen einzigen Zug gezeigt.

Wie kann etwas besser werden als sein Lehrer, wenn es *gar keinen Lehrer hat*? Genau um diese Frage tobt gerade die große Debatte: **sich selbst verbessernde KI** — Systeme, die ohne weiteren menschlichen Input immer besser werden. Heute nehmen wir sie auseinander.

Roter Faden: Wie immer gehen wir von der Schlagzeile runter bis zu den Zahlen — am Ende siehst du das komplette „Gehirn“ einer selbstlernenden KI als Tabelle aus Tensoren.

Was du hier lernst

- Warum „sich selbst Geschichten schreiben und darauf trainieren“ **nicht** funktioniert (Model Collapse)
- Die Zauberzutat echter Selbstverbesserung: der **unbestechliche Schiedsrichter**
- Die Schleife: Vorschlagen → Prüfen → Aussieben → Trainieren → von vorn
- Die zwei Ebenen: besser **können** vs. besser **gebaut** sein
- Im Mitmach-Teil: Du siehst einer KI live beim Selbst-Lernen zu — und spielst danach gegen sie

Für den Unterricht

Einstieg (5 Min): Kapitel gemeinsam lesen, Alltagsbeispiel aufgreifen.

Erkunden am Gerät (15 Min): die interaktiven Module auf herrki.de/395_selbstverbesserung ausprobieren.

Sichern mit Quiz (5 Min): „Kurzer Check" besprechen, „Das nimmst du mit" festhalten.

Der naheliegende (falsche) Plan: sich selbst Geschichten erzählen

Die Idee klingt erstmal logisch: Das Modell schreibt sich eine Million neue Übungstexte, trainiert darauf, schreibt noch bessere Texte, trainiert wieder … und schaukelt sich so nach oben. Warum klappt das nicht?

Denk an einen **Fotokopierer, der Kopien von Kopien** macht: Mit jeder Runde verschwinden feine Details, kleine Fehler werden mitkopiert und verstärken sich. Beim Modell passiert exakt das — seine Ausgaben sind ja nur eine (leicht verwaschene) Zusammenfassung dessen, was es *schon weiß*. Trainiert es darauf, lernt es seine eigenen Macken als „Wahrheit“. 2024 wurde das in der Fachzeitschrift *Nature* sauber nachgewiesen: Modelle, die über mehrere Generationen auf ihren eigenen Ausgaben trainieren, **kollabieren** — seltenes Wissen verschwindet zuerst, am Ende bleibt Einheitsbrei. Der Fachbegriff: **Model Collapse**.

Merksatz: Aus den eigenen Ausgaben allein entsteht **keine neue Information**. Wer nur sein eigenes Echo hört, hört nichts Neues.

Die Zauberzutat: ein unbestechlicher Schiedsrichter

Neue Information muss von **außen** kommen. Aber — und das ist der Punkt, den viele übersehen — „außen“ muss **kein Mensch** sein! Es reicht eine Instanz, die nicht das Modell selbst ist und ehrlich urteilt: richtig oder falsch, gewonnen oder verloren, läuft oder stürzt ab. Solche unbestechlichen Prüfer gibt es überall dort, wo die Wirklichkeit selbst die Antwort kennt:

Gebiet	Der Schiedsrichter	Sein Urteil
Spiele (Go, Schach, …)	Die Spielregeln	gewonnen / verloren / remis
Mathematik	Ein Beweisprüfer-Programm	Beweis hält / hält nicht
Programmieren	Der Computer + Tests	läuft korrekt / schlägt fehl
Naturwissenschaft	Experiment oder Simulation	Messwert stimmt / stimmt nicht

Mit so einem Schiedsrichter wird aus der Echokammer eine **Aufwärtsspirale**:

1
Vorschlagen
Das Modell erzeugt viele Lösungswege — auch verrückte

2
Prüfen
Der Schiedsrichter bewertet jeden einzelnen ehrlich

3
Aussieben
Nur bestätigte Wege werden behalten

4
Trainieren
Das Modell lernt aus dem Bestätigten

↻ … und wieder von vorn — aber jetzt schlägt das Modell schon bessere Wege vor

Genau DAS ist mit „neuen Erkenntnispfaden“ gemeint: Nicht die selbst geschriebenen Geschichten sind der Gewinn — sondern die **geprüften und bestätigten**

Lösungswege. Sie enthalten echte neue Information, denn in ihnen steckt jetzt das Urteil des Schiedsrichters, also ein Stück Wirklichkeit, das vorher nicht im Modell war.

Das Nachjustieren selbst kennst du übrigens längst: Fehler messen wie in „Fehler messen“, Gewichte anpassen wie in „Gradient Descent“. **Neu ist nur, woher das Lernsignal kommt:** nicht aus menschlichen Beispieldaten, sondern aus dem Schiedsrichter-Urteil über selbst erzeugte Versuche.

Die Belege: vier Meilensteine der Selbstverbesserung

1. AlphaGo Zero (2017) — der Urknall des Self-Play

Nur Spielregeln, kein menschliches Wissen. Das Netz spielt gegen sich selbst; der Schiedsrichter (die Regeln) sagt nach jeder Partie, wer gewonnen hat; das Netz trainiert auf seinen erfolgreichen Zügen. Ergebnis: übermenschliche Spielstärke — und Strategien, die menschliche Profis *seither von der Maschine übernehmen*. Die Schülerin wurde zur Lehrerin.

2. AlphaTensor & AlphaDev (2022/23) — echte neue Entdeckungen

Dieselbe Self-Play-Idee, aber das „Spiel“ war: *finde ein schnelleres Rechenverfahren*. Die Systeme entdeckten neue Algorithmen für Matrix-Multiplikation und Sortieren, nach denen Menschen jahrzehntelang vergeblich gesucht hatten. Schiedsrichter: mathematische Korrektheit plus Stoppuhr — beides unbestechlich. Das war der Beweis, dass die Schleife **neues Wissen** erzeugen kann, nicht nur Spielstärke.

3. FunSearch & AlphaEvolve (2023/2025) — das Sprachmodell in der Schleife

Ein Sprachmodell schreibt Programme als Lösungsvorschläge, ein Prüfprogramm bewertet sie, die besten werden weiterentwickelt — Evolution mit KI als Ideengeber. So entstanden neue mathematische Resultate. Und AlphaEvolve verbesserte sogar die Rechenkerne, mit denen Googles *nächste* Modellgeneration schneller trainiert wird: Die KI beschleunigt das Training ihrer Nachfolgerin — Selbstverbesserung über Bande.

4. Reasoning-Modelle / RLVR (2024/25) — verifizierte Denkpfade

So wurden die „Denk-Modelle“ trainiert, die du aus „Erst denken, dann reden“ kennst: Das Modell erzeugt zu Mathe- und Programmieraufgaben tausende Denkpfade — belohnt und verstärkt werden **nur die, deren Endergebnis der Schiedsrichter als richtig bestätigt** (Fachbegriff: RLVR, „Reinforcement Learning mit verifizierbaren Belohnungen“). Dabei tauchten Denk-Strategien auf, die niemand vorgegeben hatte: innehalten, zurückgehen, sich selbst korrigieren. Im berühmten DeepSeek-R1-Bericht (Januar 2025) nannten die Forscher das den „Aha-Moment“ ihres Modells.

Zwei Ebenen: besser KÖNNEN und besser GEBAUT sein

In der aktuellen Diskussion werden zwei verschiedene Dinge „Selbstverbesserung“ genannt — und es lohnt sich, sie auseinanderzuhalten:

- **Ebene 1 — das Können (die Gewichte):** Self-Play und RLVR justieren die Gewichts-Tensoren des Modells. Das Bauwerk bleibt gleich, aber es kann immer mehr. Das ist heute **Realität und Standard-Praxis** — jedes moderne Reasoning-Modell wurde so trainiert.
- **Ebene 2 — der Bauplan (Struktur und Code):** Systeme, die ihre eigene Architektur oder ihren eigenen Programmcode umbauen und die neue Version empirisch gegen die alte testen. Beispiele: automatische Architektur-Suche und 2025 die „Darwin Gödel Machine“, ein Agent, der seinen eigenen Code umschreibt und nur nachweislich bessere Versionen behält. Schiedsrichter: messbare Testergebnisse. Das ist **Forschung mit ersten echten Erfolgen** — die spektakulären Schlagzeilen träumen meist von dieser Ebene.

Beide Ebenen benutzen dieselbe Zauberzutat: Vorschläge machen, unbestechlich prüfen lassen, nur das Bestätigte behalten. Ob dabei Gewichte oder Baupläne verbessert werden, ist „nur“ eine Frage, *woran* geschraubt wird.

Runter zum Tensor: Was Self-Play mit den Gewichten macht

Die Arena gleich unten hält das Gehirn ihrer KI absichtlich winzig: eine Tabelle, die zu jeder Brettstellung **9 Zahlen** speichert — eine Bewertung je Feld. Bei AlphaGo Zero und den Reasoning-Modellen steht an genau dieser Stelle ein neuronales Netz, aber die Kette ist dieselbe: Der Spielzustand geht als **Tensor** hinein, heraus kommt ein Vektor mit einer Zahl je möglichem Zug (im großen Modell sind das Wahrscheinlichkeiten), und der Schiedsrichter liefert am Ende der Partie eine einzige Zahl — einen **Skalar**. Aus diesen drei Teilen wird ein Lernschritt: Urteil minus bisherige Erwartung ergibt den Fehler, und jede benutzte Zahl rutscht ein Stück in dessen Richtung — hier mit dem festen Faktor 0,25, im großen Modell als Gradientenschritt durch Millionen Gewichte. Mehr ist „die KI hat gelernt“ nicht: kein neuer Gedanke, nur verschobene Zahlen.

```
Brettstellung // 9 Felder: leer / X / O → Tensor [9]
Bewertung je Zug → Vektor [9] // im grossen Modell: Wahrscheinlichkeiten
Urteil des Schiedsrichters → Skalar [ ] // Sieg +1 · Remis +0,5 · Niederlage −1

// ein Lernschritt — genau der, den die Arena 6.000 Mal macht:
Zahl_neu = Zahl_alt + 0,25 × (Ergebnis − Zahl_alt)

// Model Collapse, in Tensoren gesprochen:
ohne Schiedsrichter → der Vektor wird spitz: fast alles Gewicht auf wenige Zuege,
                      die seltenen Werte rutschen gegen 0 // die Verteilung schrumpft
```

Damit ist auch der Unterschied der beiden Experimente unten reine Rechnung: In Experiment A steht im Skalar das echte Spielergebnis, in Experiment B ein Münzwurf. Gleiche Formel, gleich viele Partien — nur zeigt der eine Fehler in Richtung Wirklichkeit und der andere ins Nichts.

Deine Selbstlern-Arena: eine KI lernt Tic-Tac-Toe — ohne dich

Jetzt baust du dir die AlphaGo-Zero-Idee im Kleinen: Eine KI lernt Tic-Tac-Toe **ausschließlich durch Spiele gegen sich selbst**. Wir haben ihr NUR die Spielregeln einprogrammiert (das ist der Schiedsrichter!) und ihr ein leeres Gehirn gegeben. Kein einziger Beispielzug von einem Menschen.

Runter zum Tensor: Ihr komplettes Gehirn ist eine Tabelle: Für jede Brettstellung merkt sie sich **9 Zahlen** — „wie gut ist ein Zug auf Feld 1 bis 9?“. Nach jeder Selbst-Partie meldet der Schiedsrichter das Ergebnis, und jede benutzte Zahl wird ein Stück nachjustiert:

```
Zahl_neu = Zahl_alt + 0,25 × (Ergebnis − Zahl_alt) Ergebnis: Sieg +1 · Remis +0,5 · Niederlage −1
```

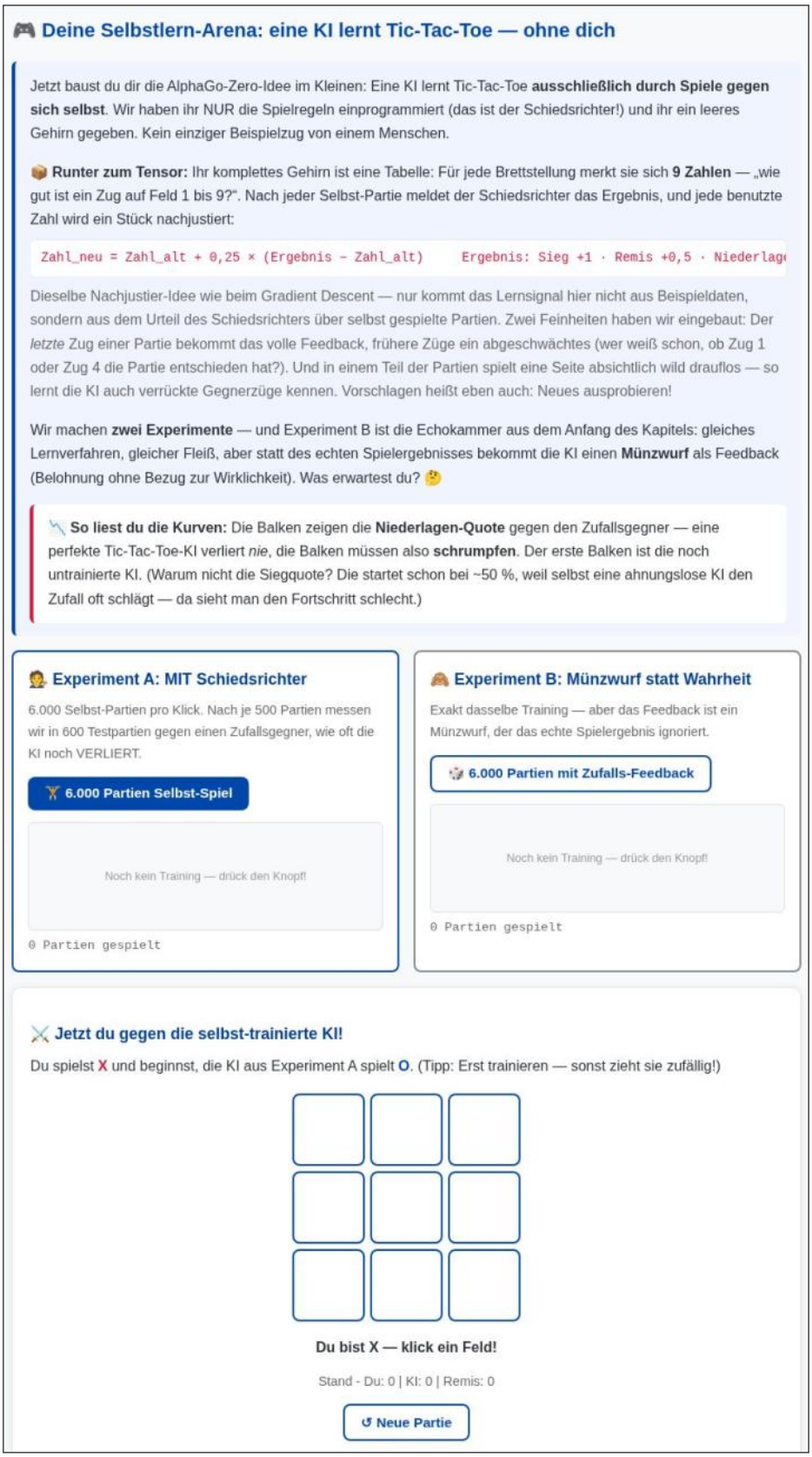
Dieselbe Nachjustier-Idee wie beim Gradient Descent — nur kommt das Lernsignal hier nicht aus Beispieldaten, sondern aus dem Urteil des Schiedsrichters über selbst gespielte Partien. Zwei Feinheiten haben wir eingebaut: Der *letzte* Zug einer Partie bekommt das volle Feedback, frühere Züge ein abgeschwächtes (wer weiß schon, ob Zug 1 oder Zug 4 die Partie entschieden hat?). Und in einem Teil der Partien spielt eine Seite absichtlich wild drauflos — so lernt die KI auch verrückte Gegnerzüge kennen. Vorschlagen heißt eben auch: Neues ausprobieren!

Wir machen **zwei Experimente** — und Experiment B ist die Echokammer aus dem Anfang des Kapitels: gleiches Lernverfahren, gleicher Fleiß, aber statt des echten Spielergebnisses bekommt die KI einen **Münzwurf** als Feedback (Belohnung ohne Bezug zur Wirklichkeit). Was erwartest du?

So liest du die Kurven: Die Balken zeigen die **Niederlagen-Quote** gegen den Zufallsgegner — eine perfekte Tic-Tac-Toe-KI verliert *nie*, die Balken müssen also **schrumpfen**. Der erste Balken ist die noch untrainierte KI. (Warum nicht die Siegquote? Die startet schon bei ~50 %, weil selbst eine ahnungslose KI den Zufall oft schlägt — da sieht man den Fortschritt schlecht.)

Am Gerät ausprobieren: herrki.de/395_selbstverbesserung#abschnitt-7

Direkt zum Modul



Interaktives Modul – am Gerät ausprobieren

Experiment A: MIT Schiedsrichter

6.000 Selbst-Partien pro Klick. Nach je 500 Partien messen wir in 600 Testpartien gegen einen Zufallsgegner, wie oft die KI noch VERLIERT.

0 Partien gespielt

Experiment B: Münzwurf statt Wahrheit

Exakt dasselbe Training — aber das Feedback ist ein Münzwurf, der das echte Spielergebnis ignoriert.

0 Partien gespielt

✓ **Siehst du den Unterschied?** Beide KIs haben gleich viele Partien „geübt“. Aber nur bei der KI mit dem ehrlichen Schiedsrichter schrumpfen die Niederlagen Richtung null — die Münzwurf-KI verliert munter weiter, egal wie lange sie „trainiert“. **Nicht das Üben macht die Selbstverbesserung, sondern das ehrliche Urteil.**

Jetzt du gegen die selbst-trainierte KI!

Du spielst **X** und beginnst, die KI aus Experiment A spielt **O**. (Tipp: Erst trainieren — sonst zieht sie zufällig!)

Du bist X — klick ein Feld!

Stand — Du: 0 · KI: 0 · Remis: 0

Der Blick ins Gehirn: die 9 Zahlen der KI

So bewertete die KI das Brett vor ihrem letzten Zug — das ist ihr *kompletter* „Gedanke“: grün = Zug erscheint gut, rot = Zug erscheint schlecht.

Noch kein Zug gemacht.

Zurück zur Anwendung

Zurück zur Schlagzeile vom Anfang — „KI verbessert sich selbst“. Du kannst diesen Satz ab jetzt zerlegen: in Vorschläge, ein Urteil und einen Lernschritt. Und du weißt, welches der drei Teile knapp ist — nicht die Ideen, sondern der unbestechliche Prüfer. Bevor du das nach oben mitnimmst, gehört die Gegenseite dazu:

Ehrlich bleiben: die Grenzen

- **Der Flaschenhals ist der Schiedsrichter, nicht der Ideen-Generator.** Für guten Schreibstil, Geschichtsdeutung oder Lebensberatung gibt es keinen unbestechlichen Prüfer — dort droht sofort wieder die Echokammer. Deshalb spielen die großen Selbstverbesserungs-Erfolge bisher fast alle in Mathe, Code und Spielen.
- **Teuer:** Die Schleife braucht Millionen von Versuchen. Deine Tic-Tac-Toe-Arena schafft das im Browser — bei großen Modellen kostet dieselbe Idee Rechenzentren.
- **Besser im Prüfbaren ≠ allgemein klüger.** Ob die Schleife jemals *voll autonom über alle Gebiete* läuft (Stichwort „Intelligenzexplosion“), ist eine offene Forschungs- und Streitfrage. Heute real: mächtige Bausteine. Nicht real: ein Perpetuum mobile der Erkenntnis.

Und damit zurück nach oben: Wenn du das nächste Mal liest „KI verbessert sich selbst ohne Menschen“, stell die eine Frage: **„Wer ist der Schiedsrichter?“** Gibt es einen unbestechlichen — Regeln, Tests, Beweise, Experimente —, ist die Meldung ernst zu nehmen. Gibt es keinen, liest du vermutlich gerade über eine sehr teure Echokammer.

Zum Nachlesen (Originalquellen)

- AlphaGo Zero: „Mastering the game of Go without human knowledge“ (Nature, 2017) (nature.com/articles/nature24270)
- Model Collapse: „AI models collapse when trained on recursively generated data“ (Nature, 2024) (nature.com/articles/s41586-024-07566-y)
- FunSearch: „Mathematical discoveries from program search with LLMs“ (Nature, 2023) (nature.com/articles/s41586-023-06924-6)
- DeepSeek-R1 („Aha-Moment“): arXiv:2501.12948 (2025) (arxiv.org/abs/2501.12948)
- STaR (Selbst-Training mit geprüften Begründungen): arXiv:2203.14465 (2022) (arxiv.org/abs/2203.14465) — das Rezept, mit dem ein Modell aus den eigenen, bestätigten Lösungswegen weiterlernt
- Darwin Gödel Machine: arXiv:2505.22954 (2025) (arxiv.org/abs/2505.22954)

Kurzer Check – mit Lösung

Lösungen für Lehrkräfte – für Lernende die Online-Version nutzen.

Frage 1: Eine KI schreibt sich selbst eine Million Übungstexte und trainiert darauf. Warum wird sie davon nicht klüger?

- Weil eine Million Texte zu wenig sind — mit einer Milliarde würde es klappen.
- **✓ In den eigenen Ausgaben steckt keine neue Information — Fehler verstärken sich sogar (Model Collapse).**
- Weil selbst geschriebene Texte grundsätzlich nicht ins Trainingsformat passen.

Fotokopie einer Fotokopie: Jede Runde verwäscht Details und verstärkt Macken. Neue Information braucht ein Urteil von AUSSEN — einen Schiedsrichter, der nicht das Modell selbst ist.

Frage 2: Was ist die entscheidende Zutat, damit Selbstverbesserung wirklich funktioniert?

- Mehr Rechenleistung.
- Ein größeres Modell mit mehr Neuronen.
- **✓ Ein unbestechlicher Schiedsrichter, der jeden Versuch ehrlich bewertet — Spielregeln, Tests, Beweise, Experimente.**

Vorschlagen können Modelle massenhaft. Der Engpass ist das ehrliche Urteil: Nur geprüfte und bestätigte Lösungswege enthalten neue Information, auf der sich sinnvoll weitertrainieren lässt.

Frage 3: In deiner Arena lernte Experiment B (Münzwurf-Feedback) nichts dazu. Warum?

- Es hat zu wenige Partien gespielt.
- **✓ Das Feedback hatte keinen Bezug zur Wirklichkeit — Belohnung ohne Wahrheit erzeugt nur Rauschen im Gehirn der KI.**
- Tic-Tac-Toe ist ohne Schiedsrichter mathematisch unlösbar.

Gleiches Lernverfahren, gleicher Fleiß, gleiche Partienzahl — nur das Urteil war wertlos. Lernen = Vorschläge + ehrliches Urteil. Fehlt das Urteil, justiert die KI ihre Zahlen in zufällige Richtungen.

Das nimmst du mit

- **Selbst erzeugte Texte enthalten keine neue Information.** Fotokopie einer Fotokopie — Fehler verstärken sich sogar (Model Collapse).
- **Der Schiedsrichter ist die Zauberzutat.** Spielregeln, Tests, Beweise, Experimente: erst ein Urteil von außen macht aus Vorschlägen echtes Lernmaterial.
- **Belohnung ohne Wahrheit ist Rauschen.** Dein Experiment B hat mit Münzwurf-Feedback gleich fleißig gespielt — und nichts gelernt.
- **Zwei Ebenen der Selbstverbesserung.** Besser *können* (die KI übt) und besser *gebaut* sein (die KI verbessert ihr eigenes Werkzeug) — das ist nicht dasselbe.
- **Der Flaschenhals ist die Prüfbarkeit.** Deshalb spielen die Erfolge in Mathe, Code und Spielen — und die Frage bei jeder Schlagzeile lautet: *Wer ist der Schiedsrichter?*

Weiter geht's: Die verifizierten Denkpfade von hier sind genau das, was in „Erst denken, dann reden“ trainiert wurde — und in „Denken im Verborgenen“ siehst du, wohin sich dieses Denken gerade weiterentwickelt. Das Nachjustieren dahinter kennst du längst aus „Gradient Descent“ — der Kreis schließt sich.

Und damit endet die Kategorie „Aktuell“. Ab hier wird es **interaktiv**: In „Trainiere auf DEINEN Daten“ arbeitest du zum ersten Mal mit deinen eigenen Zahlen — du tippst sie ein, das Modell lernt darauf.

So zitierst du dieses Kapitel

HerrKI: Die KI, die sich selbst verbessert. herrki.de/395_selbstverbesserung, ein privates Hobbyprojekt von Terragon.de, abgerufen am 09.09.2026.