Ein hervorragendes und brandaktuelles Beispiel für die **zirkuläre Validierung** findet sich in der modernen Software-Entwicklung und Datenwissenschaft. Man nennt dieses Phänomen in der Forschung auch die **„Synthetische Daten-Sackgasse“** oder die Gefahr des **„Modell-Kollapses“**.

Das Problem entsteht immer dann, wenn Menschen aus Kostengründen oder wegen Überlastung den Menschen komplett aus der Prüfschleife (Human-in-the-Loop) entfernen und eine KI die Arbeit einer anderen KI bewerten lassen.

Das Szenario: Die Entwicklung einer Betrugs-Erkennungs-KI

Stell dir eine Großbank vor, die eine neue KI entwickeln möchte, um Kreditkartenbetrug in Echtzeit zu erkennen.

- **Das Problem:** Echte Daten von Betrugsfällen sind extrem sensibel (Datenschutz/DSGVO) und selten.
- **Die „smarte“ Lösung:** Die Bank nutzt eine hochentwickelte **Generator-KI**. Diese hat die Aufgabe, Millionen von *synthetischen* (künstlichen) Kundendaten und Transaktionen zu erzeugen, die täuschend echt aussehen, aber frei erfunden sind.

1. Der Schritt zur zirkulären Validierung

Die Generator-KI spuckt vereinbarungsgemäß einen Datensatz mit 10 Millionen gefälschten Transaktionen aus. Kein menschliches Team kann 10 Millionen Zeilen Code und Daten manuell auf Logikfehler prüfen.

Also baut die Bank die zirkuläre Schleife: Sie setzt eine zweite, unabhängige **Prüf-KI** (einen „Validator“) ein. Die Aufgabe der Prüf-KI lautet: *„Geh durch diese 10 Millionen Datenpunkte und validiere, ob diese Transaktionen realistisch sind und die KI gute Arbeit geleistet hat.“*

2. Das Ergebnis: Das perfekte (und falsche) Zertifikat

Die Prüf-KI analysiert die Daten und meldet nach wenigen Minuten stolz: *„Ergebnis: 99,8 % Genauigkeit. Die Daten sind absolut realistisch, fehlerfrei und perfekt geeignet, um die neue Betrugs-Software zu trainieren.“*

Die Manager der Bank sind begeistert. Zwei KIs haben sich gegenseitig kontrolliert, das Ergebnis ist statistisch perfekt. Menschlicher Aufwand: fast null.

Warum dieses System in der Realität kollidiert

Das Problem ist die **gemeinsame statistische Basis**. Beide KIs – der Generator und der Validator – basieren letztlich auf ähnlichen mathematischen Prinzipien und wurden mit ähnlichen Mustern des Internets trainiert.

- Wenn die Generator-KI beim Erstellen der Daten einen logischen Denkfehler macht (z. B. weil sie nicht „weiß“, dass Tankstellen nachts um 3 Uhr in einem bestimmten Land keine Transaktionen über 5.000 Euro zulassen), wird sie diesen Fehler in die synthetischen Daten einbauen.
- Die Prüf-KI leidet aber unter **denselben blinden Flecken**. Da sie den gleichen Denkfehler teilt, fällt ihr die Unlogik nicht auf. Sie sieht die Daten, findet sie statistisch harmonisch und winkt sie durch.

Das böse Erwachen

Die Bank trainiert ihr finales System nun mit diesen „validierten“ Daten. Als das System live geht, wird die Bank prompt von echten Betrügern überlistet. Warum? Weil die KI mit Daten trainiert wurde, die zwar *mathematisch harmonisch* waren (weil KI 2 KI 1 recht gegeben hat), aber völlig an der komplexen, chaotischen Realität echten menschlichen Verhaltens vorbeiging.

Der Kern des Problems: Bei der zirkulären Validierung bestätigen sich zwei Systeme gegenseitig in ihren eigenen Vorurteilen und Fehlern. Es entsteht eine computationale Echo-Kammer. Der Mensch schaut von außen auf das perfekte Prüfprotokoll der zweiten KI und wiegt sich in einer völligen Scheinsicherheit.