Ein konkretes Beispiel: Medizinische Prüfungen

Moderne Sprachmodelle erzielen auf dem amerikanischen Medizinlizenz-Examen (USMLE) Ergebnisse auf Niveau erfahrener Ärzte. Das ist ein echter Benchmark, der jahrzehntelang als valider Nachweis medizinischer Kompetenz galt.

Stellt man dasselbe Modell dann vor echte klinische Situationen – unvollständige Patientenakten, widersprüchliche Symptome, die Notwendigkeit nachzufragen – treten systematische Fehler auf, die einem ausgebildeten Arzt nicht passieren würden. Das Modell hat gelernt, gut auf dem Test abzuschneiden, ohne die zugrunde liegende Kompetenz zu erwerben, die der Test ursprünglich messen sollte.

Der Benchmark sagt: *„Arzt-Niveau."* Die Realität sagt: *„Etwas anderes."*

Der strukturelle Unterschied zum vorigen Punkt

	Systematische Fehlkalibrierung	**Benchmark-Entkopplung**
Wo liegt der Fehler?	In der Wahrnehmung des Menschen	Im Messinstrument selbst
Wer/was irrt sich?	Die Person, die das KI-System beurteilt	Der Benchmark als Proxy für Fähigkeit
Mechanismus	Oberflächenmerkmale täuschen die Intuition	Das System optimiert den Test, nicht die dahinterliegende Fähigkeit

Mit anderen Worten:

- Bei der **Fehlkalibrierung** täuscht das menschliche Auge sich selbst – weil Rilke-Gedichte nach tiefer Sprachkompetenz *aussehen*.
- Bei der **Benchmark-Entkopplung** streckt sich das Maßband – das Messinstrument war mal valide, verliert aber seine Aussagekraft, weil Systeme auf den Test hin optimiert werden, nicht auf das, was er messen soll.

Das ist Goodharts Gesetz in neuer Form: *Sobald eine Metrik zum Ziel wird, hört sie auf, eine gute Metrik zu sein.*

Warum das gravierender ist

Bei der Fehlkalibrierung kann ein aufmerksamer Mensch seinen Irrtum korrigieren – wenn er merkt, dass Oberflächenmerkmale ihn täuschen, passt er sein Urteil an.

Bei der Benchmark-Entkopplung ist das schwerer. Der Fehler steckt im Messwerkzeug selbst, dem man oft vertraut, gerade *weil* es formalisiert und scheinbar objektiv ist. Man bemerkt das Problem oft erst, wenn das System in der Praxis eingesetzt wird und scheitert – also zu spät.

Die beiden Phänomene können sich auch gegenseitig verstärken: Wenn Benchmarks unzuverlässig werden, greifen Menschen noch stärker auf oberflächliche Eindrücke zurück – und die Fehlkalibrierung nimmt zu.