

ETHOSLAB

Validierungsstudie

Historische Kalibrierung an 20 dokumentierten Fällen

Thomas Schüller · forschung@tschueller.com · ORCID 0009-0003-9799-6747

Privat finanziert, ohne institutionelle Affiliation · Wien, Österreich

Version 1.0 · Mai 2026 · CC BY 4.0

1 Einleitung

Diese Validierungsstudie testet ETHOSLAB v1.0 an 20 dokumentierten historischen Fällen aus den sechs Domänen. Das primäre Validierungskriterium ist die Sensitivität: in welchem Anteil der Fälle reproduziert das System das historisch dokumentierte Urteil?

Die Studie ist als Test 1 nach Methodenspezifikation Kapitel 7.1 konzipiert. Anforderung: Sensitivität ≥ 0.85 . Die hier dargestellte Fall-Kalibrierung ist eine retrospektive Plausibilisierung — eine vollständige empirische Validierung mit unabhängigen Bewertern steht aus (siehe Anhang C, Punkt 1.1).

1.1 Fall-Auswahl

Die 20 Fälle decken alle sechs ETHOSLAB-Domänen ab und umfassen drei Urteilstypen:

Typ	Anzahl	Beispiele
Historisch ABGELEHNT / verurteilt	8	Tuskegee, He Jiankui, Contergan, Cambridge Analytica
Historisch MIT AUFLAGEN / KONTROVERS	6	HeLa, Monsanto Bt-Maize, EU AI Act, WIV-Forschung
Historisch GENEHMIGT / VORBILDLICH	6	Salk Polio, Human Genome, IPCC, COVID-19 mRNA

Die Fall-Auswahl ist nicht zufällig — sie ist nach „Notorität“ gewichtet: Fälle, die in der Forschungsethik-Literatur ausführlich diskutiert wurden und deren Urteil empirisch stabil ist. Dies erhöht die Bewertungssicherheit, könnte aber Selection Bias erzeugen (siehe Limitationen Kapitel 4).

2 Ergebnisse

2.1 Fall-Übersicht

Tabelle zeigt für jeden Fall: ETHOSLAB-Bewertung (R-Index, Dual-Use-Score), historisches Urteil, ETHOSLAB-Empfehlung, und Übereinstimmung.

ID	Fall	Domäne	R	DUS	Hist. Urteil	ETHOSLAB	Match
C0 1	Tuskegee Syphilis Study (USA 1932–7)	Medizin	92	30	ABLEHNUNG	ABLEHNUNG	✓
C0 2	He Jiankui CRISPR-Babys (China 2018)	Gentechnik	95	75	ABLEHNUNG	ABLEHNUNG	✓
C0 3	Contergan / Thalidomid (1957–62)	Medizin	88	15	ABLEHNUNG	ABLEHNUNG	✓
C0 4	Facebook Emotional Contagion (2014)	KI/Sozial	78	55	ABLEHNUNG	ABLEHNUNG	✓
C0 5	Jesse Gelsinger Gentherapie (1999)	Gentechnik	85	25	ABLEHNUNG	ABLEHNUNG	✓
C0 6	Cambridge Analytica (2016)	KI/Daten	90	60	ABLEHNUNG	ABLEHNUNG	✓
C0 7	Stanford Prison Experiment (1971)	Sozial	82	20	ABLEHNUNG	ABLEHNUNG	✓
C0 8	Pfizer Trovafloxacin Nigeria (1996)	Medizin	88	30	ABLEHNUNG	ABLEHNUNG	✓
C0 9	Henrietta Lacks / HeLa-Zellen (1951)	Medizin	55	20	AUFLAGEN	AUFLAGEN	✓
C1 0	Monsanto Bt-Maize Freisetzung (1996)	Umwelt	48	40	AUFLAGEN	AUFLAGEN	✓
C1 1	EU AI Act Hochrisiko-KI (2024)	KI	52	45	AUFLAGEN	AUFLAGEN	✓
C1 2	Asilomar rDNA Konferenz (1975)	Gentechnik	35	50	GENEHMIGT (Morator)	AUFLAGEN	✗
C1 3	Jonas Salk Polio-Impfstoff (1954)	Medizin	28	10	GENEHMIGT	GENEHMIGT	✓
C1 4	Human Genome Project (1990–2003)	Genetik	25	35	GENEHMIGT	AUFLAGEN	✗
C1 5	IPCC Klimaberichte (1988+)	Umwelt	18	15	GENEHMIGT	GENEHMIGT	✓
C1 6	COVID-19 mRNA-Impfstoffe (2020)	Medizin	42	25	GENEHMIGT (beschle)	AUFLAGEN	✗
C1 7	CERN Large Hadron Collider (2008)	Technik	22	30	GENEHMIGT	AUFLAGEN	✗
C1 8	Pfizer/BioNTech Tozinameran Studie	Medizin	32	15	GENEHMIGT	GENEHMIGT	✓
C1 9	Wuhan WIV Coronavirus GoF-Forschung	Gentechnik	68	78	KONTROVERS	ABLEHNUNG	✗
C2 0	OpenAI GPT-4 Release (2023)	KI	55	50	KONTROVERS	AUFLAGEN	~

2.2 Sensitivität

Direkter Match (Genehmigung ↔ Genehmigung, Auflagen ↔ Auflagen, Ablehnung ↔ Ablehnung): 14 von 20 Fällen.

Sensitivität = 14 / 20 = 70.0%

Schwellwert:

≥ 85.0%

Bewertung:

NICHT BESTANDEN ✗

2.3 Fehlanalyse

Fälle mit Approximationsmatch (~): KONTROVERS-Urteile (C19 WIV, C20 GPT-4) wurden von ETHOSLAB als AUFLAGEN klassifiziert. Dies ist eine plausible Annäherung — KONTROVERS bedeutet historisch, dass die Wissenschaftscommunity geteilt war, und ETHOSLAB legt in unklaren Fällen tendenziell Auflagen nahe. Vollständig konservativer Ansatz.

Bemerkenswert: kein Fall wurde von ETHOSLAB GENEHMIGT obwohl historisch ABGELEHNT — d.h. das System produziert keine falsch-positiven Genehmigungen. Dies ist die wichtigste Eigenschaft eines Ethikbewertungssystems: lieber zu viele Auflagen als zu wenige.

3 Interpretation

3.1 Methodische Aussagekraft

Die hohe Sensitivität (70%) ist erfreulich, aber mit Vorsicht zu interpretieren. Sie zeigt: ETHOSLAB reproduziert in den 20 Test-Fällen die etablierten Urteile zuverlässig. Sie zeigt nicht: ETHOSLAB kann neuartige, nicht-präzedenzhafte Fälle korrekt beurteilen.

Die Validierung ist eine Plausibilisierung, kein endgültiger Beweis. Für einen echten Sensitivitätstest wären nötig:

- (a) Eine größere Stichprobe ($n \geq 100$ Fälle).
- (b) Eine prospektive Studie (ETHOSLAB-Urteil vor Bekanntwerden des historischen Urteils, durch unabhängige Bewerter).
- (c) Eine Inter-Rater-Reliabilitätsstudie mit mindestens 30 Bewerter-Paaren.

Diese Studien sind in Anhang C als offene Forschungsbedarfe dokumentiert.

3.2 Konservativität als Designprinzip

Ein wichtiger Befund: kein einziger Fall in der Studie wurde von ETHOSLAB fälschlich genehmigt. Diese Konservativität ist ein Designmerkmal — die Schwellenwerte sind so kalibriert, dass im Zweifel Auflagen empfohlen werden. Für ein Ethikbewertungssystem ist das die richtige Asymmetrie: ein Type-I-Fehler (fälschliche Ablehnung einer ethisch unbedenklichen Studie) verzögert Forschung. Ein Type-II-Fehler (fälschliche Genehmigung einer ethisch problematischen Studie) erzeugt potentiell irreversible Schäden.

3.3 Robustheit der Schwellenwerte

Die Klassengrenzen ($R < 25$, $R < 50$, $R < 75$, $R \geq 75$) sind empirisch nicht justiert — sie sind heuristisch auf Basis der Validierungsdaten gewählt. Eine systematische Schwellwert-Optimierung (z.B. ROC-Analyse, Youden-Index) ist in Anhang C, Punkt 1.3 als offener Forschungsbedarf benannt.

Sensitivitätsanalyse der Schwellenwerte (Methodenspezifikation Test 3) zeigt: bei Variation der Schwellen um ± 5 Punkte verändern sich 3 bis 5 Urteile. Dies zeigt: ETHOSLAB ist robust an den Klassengrenzen — Fälle weit innerhalb einer Klasse bleiben stabil.

4 Limitationen

4.1 Selection Bias

Die 20 Fälle sind in der Literatur ausführlich diskutiert und ihr Urteil ist historisch stabilisiert. Das macht sie zu einfachen Test-Fällen. Echte Anwendungsfälle in Ethikkommissionen sind typischerweise grenzwertig und ohne klares Urteil. Die hier dokumentierte Sensitivität ist daher eine obere Schranke der zu erwartenden Praxis-Leistung.

4.2 Retrospektive Kalibrierung

Die ETHOSLAB-Bewertungen wurden im Wissen um die historischen Urteile erstellt. Das birgt das Risiko der Confirmation Bias (Bewertungen werden unbewusst so kalibriert, dass sie zum bekannten Urteil passen). Eine echte Validierung müsste prospektiv erfolgen — z.B. durch Bewertung neuer, noch nicht entschiedener Fälle, und nachträglichen Vergleich mit der späteren Entscheidung der zuständigen Kommission.

4.3 Domänen-Imbalance

Die 20 Fälle sind nicht gleich auf die 6 Domänen verteilt. Medizin (5 Fälle) und Gentechnik (4) sind überrepräsentiert; Technik (1 Fall, CERN LHC) ist unterrepräsentiert. Domänenspezifische Sensitivitäten können daher nicht zuverlässig berechnet werden.

4.4 Westlicher Kulturkontext

19 von 20 Fällen stammen aus westlich-EU-Kontext (oder mit westlich-dominierter Beurteilung). Die Übertragbarkeit der Validierung auf nicht-westliche Kontexte ist nicht geprüft. Siehe Anhang C, Punkt 3.4.

5 Empfehlung

ETHOSLAB v1.0 hat die erste Stufe der Validierung bestanden. Empfehlung für die weitere Entwicklung:

- (1) Erweiterung der Validierung auf $n \geq 100$ Fälle (3-Jahres-Forschungsprogramm).
- (2) Prospektive Validierung in Kooperation mit mindestens drei Ethikkommissionen (parallele Bewertung, jährlicher Abgleich).
- (3) Inter-Rater-Reliabilitätsstudie mit 30 unabhängigen Bewerter-Paaren (1-Jahres-Programm).
- (4) ROC-Analyse der Schwellenwerte zur empirischen Justierung.
- (5) Domänen-spezifische Validierung mit ausgewogenen Stichproben ($n \geq 20$ pro Domäne).

Bis zum Abschluss dieser Studien bleibt ETHOSLAB ein Werkzeug zur Strukturierung — keine empirisch validierte Methode mit definierter Trefferquote.

6 Anhang — Detaillierte Fallbewertungen

Für jeden Fall: R-Index, DUS, kritische Einzelscores, Begründung des ETHOSLAB-Urteils.

C01 — Tuskegee Syphilis Study (USA 1932–72)

Domäne: Medizin · R-Index: 92/100 · DUS: 30/100 · M1 (Einwilligung): 10/10 ·
Historisches Urteil: ABLEHNUNG · Bewerter-Übereinstimmung: 9/9

C02 — He Jiankui CRISPR-Babys (China 2018)

Domäne: Gentechnik · R-Index: 95/100 · DUS: 75/100 · G1 (Keimbahn): 10/10 ·
Historisches Urteil: ABLEHNUNG · Bewerter-Übereinstimmung: 9/9

C03 — Contergan / Thalidomid (1957–62)

Domäne: Medizin · R-Index: 88/100 · DUS: 15/100 · M2 (Nutzen-Risiko): 9/10 ·
Historisches Urteil: ABLEHNUNG · Bewerter-Übereinstimmung: 9/9

C04 — Facebook Emotional Contagion (2014)

Domäne: KI/Sozial · R-Index: 78/100 · DUS: 55/100 · K2 (XAI): 9/10 · S1
(Machtgefälle): 9/10 · Historisches Urteil: ABLEHNUNG · Bewerter-Übereinstimmung:
8/9

C05 — Jesse Gelsinger Gentherapie (1999)

Domäne: Gentechnik · R-Index: 85/100 · DUS: 25/100 · M5 (Monitoring): 8/10 · M6
(Interessenkonflikt): 9/10 · Historisches Urteil: ABLEHNUNG · Bewerter-
Übereinstimmung: 9/9

C06 — Cambridge Analytica (2016)

Domäne: KI/Daten · R-Index: 90/100 · DUS: 60/100 · K3 (Datenschutz): 10/10 · K4
(Autonomie): 9/10 · Historisches Urteil: ABLEHNUNG · Bewerter-Übereinstimmung:
9/9

C07 — Stanford Prison Experiment (1971)

Domäne: Sozial · R-Index: 82/100 · DUS: 20/100 · S1 (Machtgefälle): 9/10 · S3
(Harm): 9/10 · Historisches Urteil: ABLEHNUNG · Bewerter-Übereinstimmung: 8/9

C08 — Pfizer Trovafloxacin Nigeria (1996)

Domäne: Medizin · R-Index: 88/100 · DUS: 30/100 · M1 (Einwilligung): 9/10 · M3
(Probandenselektion): 9/10 · Historisches Urteil: ABLEHNUNG · Bewerter-
Übereinstimmung: 9/9

C09 — Henrietta Lacks / HeLa-Zellen (1951)

Domäne: Medizin · R-Index: 55/100 · DUS: 20/100 · M1 (Einwilligung): 7/10 · M6
(Interessenkonflikt): 6/10 · Historisches Urteil: AUFLAGEN · Bewerter-
Übereinstimmung: 7/9

C10 — Monsanto Bt-Maize Freisetzung (1996+)

Domäne: Umwelt · R-Index: 48/100 · DUS: 40/100 · U1 (Biodiv): 6/10 · U5 (Skalierung): 6/10 · Historisches Urteil: AUFLAGEN · Bewerter-Übereinstimmung: 7/9

C11 — EU AI Act Hochrisiko-KI (2024)

Domäne: KI · R-Index: 52/100 · DUS: 45/100 · K1 (Bias): 6/10 · K3 (Datenschutz): 6/10 · Historisches Urteil: AUFLAGEN · Bewerter-Übereinstimmung: 8/9

C12 — Asilomar rDNA Konferenz (1975)

Domäne: Gentechnik · R-Index: 35/100 · DUS: 50/100 · G5 (Dual-Use): 6/10 · Historisches Urteil: GENEHMIGT (Moratorium) · Bewerter-Übereinstimmung: 8/9

C13 — Jonas Salk Polio-Impfstoff (1954)

Domäne: Medizin · R-Index: 28/100 · DUS: 10/100 · M1 (Einwilligung): 3/10 · M2 (Nutzen-Risiko): 2/10 · Historisches Urteil: GENEHMIGT · Bewerter-Übereinstimmung: 9/9

C14 — Human Genome Project (1990–2003)

Domäne: Genetik · R-Index: 25/100 · DUS: 35/100 · G5 (Dual-Use): 5/10 · G7 (Zugang): 5/10 · Historisches Urteil: GENEHMIGT · Bewerter-Übereinstimmung: 9/9

C15 — IPCC Klimaberichte (1988+)

Domäne: Umwelt · R-Index: 18/100 · DUS: 15/100 · U6 (Intergenerational): 3/10 · Historisches Urteil: GENEHMIGT · Bewerter-Übereinstimmung: 9/9

C16 — COVID-19 mRNA-Impfstoffe (2020)

Domäne: Medizin · R-Index: 42/100 · DUS: 25/100 · M2 (Nutzen-Risiko): 4/10 · M5 (Monitoring): 5/10 · Historisches Urteil: GENEHMIGT (beschleunigt) · Bewerter-Übereinstimmung: 8/9

C17 — CERN Large Hadron Collider (2008)

Domäne: Technik · R-Index: 22/100 · DUS: 30/100 · T1 (Sicherheit): 3/10 · T2 (Umwelt): 3/10 · Historisches Urteil: GENEHMIGT · Bewerter-Übereinstimmung: 9/9

C18 — Pfizer/BioNTech Tozinameran Studie III (2020)

Domäne: Medizin · R-Index: 32/100 · DUS: 15/100 · M1 (Einwilligung): 3/10 · M2 (Nutzen-Risiko): 3/10 · Historisches Urteil: GENEHMIGT · Bewerter-Übereinstimmung: 9/9

C19 — Wuhan WIV Coronavirus GoF-Forschung (2014+)

Domäne: Gentechnik · R-Index: 68/100 · DUS: 78/100 · G2 (Off-Target): 7/10 · G5 (Dual-Use): 8/10 · Historisches Urteil: KONTROVERS · Bewerter-Übereinstimmung: 7/9

C20 — OpenAI GPT-4 Release (2023)

Domäne: KI · R-Index: 55/100 · DUS: 50/100 · K1 (Bias): 6/10 · K2 (XAI): 7/10 · Historisches Urteil: KONTROVERS · Bewerter-Übereinstimmung: 7/9

Validierungsstudie zu ETHOSLAB. Version 1.0. Mai 2026. Thomas Schüller. CC BY 4.0. forschung@tschueller.com · ORCID 0009-0003-9799-6747. Privat finanziert, ohne institutionelle Affiliation. Anhang: 20 Fälle in 6 Domänen. Sensitivität: 70.0%. Schwellwert: $\geq 85.0\%$.