

HEALTH.AI ELI TOOLBOX / WORKSHOP 03

Ethik, Verantwortung und die Rolle von Bias in KI-Systemen

4 STUNDEN

Einführung · Ablaufplan · Methoden · Material

Konzeption und Realisierung: K8 Institut für strategische Ästhetik gGmbH

Einführung & Relevanz

Ethik, Verantwortung und die Rolle von Bias in KI-Systemen



> *Wir wissen, dass unsere KI nicht alle Patient:innengruppen gleich gut behandelt – aber wir wissen nicht genau, warum, und wie wir das ethisch bewerten sollen.*

Workshopinhalte

- Grundlagen der Medizinethik: Autonomie, Nicht-Schaden, Wohltun und Gerechtigkeit
- Ethische Prinzipien für KI: Vertrauenswürdigkeit, Transparenz, Fairness und Nachvollziehbarkeit
- Bias, Diskriminierung und Ungleichheit in Trainingsdaten und Modellen
- Fallstudien zu Diskriminierung in Diagnosesystemen oder Triage-Algorithmen
- Reflexion ethischer Verantwortung im Team
- Vertrauen und Vertrauenswürdigkeit: Kompetenz/Fähigkeit, Integrität und Wohlwollen

Learning Outcome

Die Teilnehmenden verstehen, wie Biases entstehen, wie Diskriminierung vermieden werden kann und wie ethische Prinzipien in die Gestaltung und Bewertung von KI-Systemen integriert werden.

Konzeption und Realisierung: K8 Institut für strategische Ästhetik gGmbH

Ablaufplan

Vorschlag eines Ablaufs

Dauer: 4 Arbeitsstunden (240 Minuten) · Beispielhafter Start: 09:00 Uhr (frei verschiebbar)

Phase	Zeit	Methode / Format	Inhalt & Ziel	Material / Quelle
1. Einstieg / Sensibilisierung (30 Min.)	09:00–09:05 (5 Min.)	Begrüßung & Rahmung (Plenum)	– Ziele, Ablauf und Rollen im Workshop vorstellen – Spielregel: Fallbeispiele werden vertraulich behandelt; der Workshop ersetzt keine förmliche Ethikvotum-Prüfung	kein Material für diesen Baustein
	09:05–09:15 (10 Min.)	Impromptu Networking (Liberating Structures) Methodenbeschreibung	– 2 kurze Runden im Zweiergespräch, Partner:in wechselt – Leitfragen: „Was bedeutet für dich ‚vertrauenswürdige KI'?" / „Wo hast du selbst erlebt oder gehört, dass eine Diagnose, ein Medikament oder ein Gerät bei Frauen und Männern unterschiedlich gut funktioniert hat?" – Kernpunkte kurz an Pinnwand/Whiteboard sammeln	kein Material nötig – aktiviert Vorwissen & persönliche Betroffenheit
	09:15–09:30 (15 Min.)	Kurzimpuls „Medizinethik trifft KI-Ethik" (Plenum)	– Die 4 medizinethischen Prinzipien (Autonomie, Nicht-Schaden, Wohltun, Gerechtigkeit) den 4 KI-Ethik-Prinzipien (Vertrauenswürdigkeit, Transparenz, Fairness, Nachvollziehbarkeit) gegenüberstellen – Das Ethikvotum als institutionelle Absicherung medizinethischer Prinzipien – Einordnung des Ethikantragsverfahrens – Ausblick auf die Themen des Tages: Bias, Verantwortung, Fairness by Design	Whitepaper „Anforderungen für Ethikanträge von Forschungsvorhaben"
2. Vertiefung / intensive thematische Auseinandersetzung (105 Min.)	09:30–09:35 (5 Min.)	Einführung Lernstationen (Plenum)	– Ablauf & Rollen für die Stationsarbeit erklären, Materialien austeilen	—

Konzeption und Realisierung: K8 Institut für strategische Ästhetik gGmbH

Phase	Zeit	Methode / Format	Inhalt & Ziel	Material / Quelle
2. Vertiefung / intensive thematische Auseinandersetzung (105 Min.)	09:35–10:10 (25 Min.)	Shift & Share (Liberating Structures, adaptiert) Insel A Methodenbeschreibung	Insel A – Gender Health Data Gap: Wenn fehlerhafte Daten die Gesundheit gefährden. 10 Min. stille Einzel-Lektüre der Beispiel-Sammlung + 15 Min. gemeinsame Diskussion. – KI/Gesichtserkennung: Fehlerraten von bis zu 34,7 % bei dunkelhäutigen Frauen gegenüber 0,8 % bei hellhäutigen Männern in kommerziellen Systemen (Buolamwini & Gebru 2018; vertieft in Buolamwini, „Unmasking AI", 2023) – Arzneimitteldosierung: Die Zolpidem-Dosis (Ambien) musste 2013 für Frauen halbiert werden, da Zulassungsstudien überwiegend an Männern erfolgten (FDA 2013; Zhao et al. 2023) – Diagnostik: Herzinfarkt-Symptome bei Frauen weichen häufig vom „klassischen", an Männern erforschten Muster ab (Gesundheitsforschung BMFTR) – Aktuell aus der Forschungsförderung: BMFTR-Initiative „Reduzierung des Gender Data Gap in der klinischen Forschung" (2025, 43 Projekte, rd. 5,8 Mio. €) – ein Projekt untersucht gezielt die Geschlechtergerechtigkeit eines Deep-Learning-Modells zur Sepsisdiagnostik – Vertiefende Lektüre als Diskussionsgrundlage: Criado-Perez, „Unsichtbare Frauen" (2020); Endler, „Das Patriarchat der Dinge" (2021) Leitfrage für die Gruppe: Welche dieser Muster könnten sich in unseren eigenen Daten/Modellen in unseren Projekten wiederholen?	Buolamwini & Gebru (2018): Gender Shades Buolamwini (2023): Unmasking AI FDA (2013): Safety Communication Zolpidem BMFTR: Reduzierung des Gender Data Gap weitere Literatur: siehe Verzeichnis am Ende
	10:10–10:35 (25 Min.)	Shift & Share (Liberating Structures, adaptiert) Insel B Methodenbeschreibung	Insel B – Verantwortung & Vertrauenswürdige KI: 10 Min. stille Einzel-Lektüre + 15 Min. gemeinsame Diskussion. – Wer trägt Verantwortung, wenn ein KI-System diskriminierende Ergebnisse liefert – Entwickler:in, Betreiber:in oder beide gemeinsam? Abgrenzung „gemeinsame Verantwortlichkeit" von anderen Verarbeiterrollen – Cybersicherheit als Teil vertrauenswürdiger KI: Anforderungen der KI-Verordnung an Hochrisiko-KI-Systeme, Zusammenhang von Robustheit und Fairness Leitfrage für die Gruppe: Wo ist die Verantwortung für Bias-Prüfung in unserem Projekt heute nicht eindeutig geregelt?	Orientierungshilfe „Gemeinsame Verantwortlichkeit" Cybersicherheit für Hochrisiko-KI-Systeme

Phase	Zeit	Methode / Format	Inhalt & Ziel	Material / Quelle
2. Vertiefung / intensive thematische Auseinandersetzung (105 Min.)	10:35–10:55 (25 Min.)	Shift & Share (Liberating Structures, adaptiert) Insel C Methodenbeschreibung	Insel C – Fair & inklusiv gestalten: 10 Min. stille Einzel-Lektüre + 15 Min. gemeinsame Diskussion. – Privacy by Design (Art. 25 DSGVO) als struktureller Hebel: Datenminimierung und Voreinstellungen beeinflussen auch die Repräsentativität von Trainingsdaten – Accessibility by Design (Barrierefreiheitsstärkungsgesetz): Barrierefreiheit als eigene Gerechtigkeitsdimension – wer wird bei der Produktgestaltung mitgedacht, wer nicht? Leitfrage für die Gruppe: Wo überschneiden sich Datenschutz-by-Design und Fairness-by-Design in unserem Projekt?	Checkliste „Privacy by Design“ OnePager „Accessibility by Design“
	10:55–11:15 (20 Min.)	1-2-4-All (Liberating Structures) Methodenbeschreibung	– Einzeln → Zweier → Vierer → Plenum: Was war neu? Was bleibt unklar? – Zusammenführung aller drei Inseln, Leitfrage: Wo verstärken sich Datenlücken (A), unklare Verantwortung (B) und fehlende Fairness im Design (C) gegenseitig – und was ist unser wichtigster gemeinsamer Hebel?	kein neues Material – Synthese der Lernstationen-Inhalte
Pause	11:15–11:30 (15 Min.)	Pause	—	—
3. Diskussion / Austausch / Diskurs – Arbeiten mit dem Erfahrenen (65 Min.)	11:30–11:35 (5 Min.)	Rahmung durch Expert:in (Plenum)	– Kurze Vorstellung der/des Erfahrenen (Ethik- und/oder KI-Bias-Praxis) – Kurze Erinnerung: Der Austausch ersetzt keine förmliche Ethikvotum-Prüfung	kein Material für diesen Baustein
	11:35–12:15 (40 Min.)	Troika Consulting (Liberating Structures) Beratungsrunden Methodenbeschreibung	– Ausgangspunkt: die Fallstudie Obermeyer et al. (2019, Science) – ein in den USA breit eingesetzter Risikoalgorithmus unterschätzte systematisch den Versorgungsbedarf Schwarzer Patient:innen, weil er Gesundheitskosten statt tatsächlichem Bedarf als Proxy nutzte – Dreiergruppen (Trios), Rollen wechseln reihum: Übertragung auf eigene Praxisfälle (z. B. Repräsentativität der eigenen Trainingsdaten, Triage-/Diagnosemodelle) – Erfahrene:r steht während der Runden für Rückfragen aus allen Trios bereit – Checkliste Privacy by Design & Orientierungshilfe Gemeinsame Verantwortlichkeit als Prüfraster nutzen	Obermeyer et al. (2019): Dissecting racial bias erneut: Checkliste „Privacy by Design“ erneut: Orientierungshilfe „Gemeinsame Verantwortlichkeit“

Phase	Zeit	Methode / Format	Inhalt & Ziel	Material / Quelle
3. Diskussion / Austausch / Diskurs – Arbeiten mit dem Erfahrenen (65 Min.)	12:15–12:35 (20 Min.)	Troika Consulting Plenums-Austausch	– 2–3 besonders relevante Fälle im Plenum mit der/dem Erfahrenen vertiefen, Rückbezug auf den Obermeyer-Fall als gemeinsame Referenz – Reflexion: Welche ethische Verantwortung tragen wir als Team?	kein zusätzliches Material
4. Sicherung / Lernergebnis = Ausklang (25 Min.)	12:35–12:55 (20 Min.)	What, So What, Now What? (W^3) (Liberating Structures) Methodenbeschreibung	– What? Was habe ich heute zu Bias, Gender Health Data Gap und Vertrauenswürdigkeit mitgenommen? (kurze Einzelnotiz, dann Kleingruppe) – So what? Was bedeutet das für mein Projekt/mein Team? – Now what? Was setze ich als Nächstes um? – Festhalten auf Canvas „Mein Ethik-Kompass“ (Was ich weiß / Wo ich nachfragen muss / Wen ich kontaktiere)	Canvas „Mein Ethik-Kompass“ (workshopeigenes Arbeitsblatt)
	12:55–13:00 (5 Min.)	Abschluss & Verabschiedung (Plenum)	– Dank an alle Beteiligten, insbesondere die/den Erfahrene:n – Hinweis auf alle Materialien & die Literaturliste als Handout zum Nachlesen	gesamter Materialien- und Literaturkatalog – siehe Verzeichnisse am Ende

Methoden

Methoden & Formate

5 Formate für einen abwechslungsreichen und praxisorientierten Workshop.

Impromptu Networking

Kurze Gespräche in wechselnden Zweiergruppen, um persönliche Erfahrungen und Fragen zu ethischen Herausforderungen von KI sichtbar zu machen.

1-2-4-All

Ethische Fragestellungen zunächst einzeln, anschließend zu zweit, zu viert und schließlich in der gesamten Gruppe reflektieren.

What, So What, Now What? (W^3)

Erkenntnisse aus der Auseinandersetzung mit Bias und Fairness reflektieren und daraus konkrete Handlungsschritte ableiten.

Shift & Share

Verschiedene Perspektiven auf Bias, Fairness und Vertrauenswürdigkeit werden an Themenstationen vorgestellt und gemeinsam diskutiert.

Troika Consulting

Kollegiale Beratung in Dreiergruppen, um konkrete ethische Herausforderungen aus der KI-Forschung gemeinsam zu bearbeiten.

Material

Quellen & Arbeitsmaterialien

Die Materialien sind als Text aufgeführt. Die Onlinefassung der ELI-Toolbox enthält gegebenenfalls aktualisierte Zugänge.

Cybersicherheit für Hochrisiko-KI-Systeme

Checkliste „Privacy by Design" (Art. 25 DSGVO)

OnePager „Accessibility by Design: Barrierefreiheit von Produkten" (BFSG)

Criado-Perez, C. (2020). Unsichtbare Frauen: Wie eine von Daten beherrschte Welt die Hälfte der Bevölkerung ignoriert. München: Penguin/btb.

Obermeyer, Z., Powers, B., Vogeli, C. & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447–453.

Orientierungshilfe „Gemeinsame Verantwortlichkeit" (Bayerischer Landesbeauftragter für den Datenschutz)

Whitepaper „Anforderungen für Ethikanträge von Forschungsvorhaben" (§ 15 BO-Ä Saarland)

Buolamwini, J. (2023). Unmasking AI: My Mission to Protect What Is Human in a World of Machines. New York: Random House.

Endler, R. (2021). Das Patriarchat der Dinge: Warum die Welt Frauen nicht passt. Reinbek: Rowohlt.

OnePager: „Accessibility by Design: Barrierefreiheit von Produkten": Ein Kurzüberblick über das Barrierefreiheitsstärkungsgesetz (BFSG).

Konzeption und Realisierung: K8 Institut für strategische Ästhetik gGmbH

Material

Weiterführende Materialbeschreibungen

Erläuterungen zu den ergänzenden Quellen und Arbeitsmaterialien.

Buolamwini, J. & Gebru, T. (2018). Gender Shades

Intersectional Accuracy Disparities in Commercial Gender Classification. Proceedings of Machine Learning Research, 81, 1–15.

U.S. Food and Drug Administration (2013). Drug Safety Communication

FDA requires lower recommended doses for certain drugs containing zolpidem.

Zhao, H. et al. (2023). Making a „sex-difference fact“

Ambien dosing at the interface of policy, regulation, women's health, and biology. History of the Human Sciences.

Bose, D., Seguí-Gómez, M. & Crandall, J. R. (2011). Vulnerability of Female Drivers Involved in Motor Vehicle Crashes

An Analysis of US Population at Risk. American Journal of Public Health, 101(12), 2368–2373.

Bundesministerium für Forschung, Technologie und Raumfahrt / Gesundheitsforschung BMFTR

Reduzierung des Gender Data Gap in der klinischen Forschung (Förderinitiative 2025; 43 Projekte, rd. 5,8 Mio. €, inkl. Projekt zur Geschlechtergerechtigkeit eines Sepsis-KI-Modells).

Cybersicherheit für Hochrisiko-KI-Systeme“

Nach der KI-Verordnung sind Anbieter von Hochrisiko-KI-Systemen verpflichtet, Cybersicherheit zu gewährleisten. Dieser Beitrag erläutert die technischen Grundzüge von cybersicherer KI sowie die gesetzlichen Anforderungen der KI-Verordnung. Darauf aufbauend werden Maßnahmen zur praktischen Umsetzung der Anforderungen und zur Vertragsgestaltung vorgestellt. Abschließend wird aufgezeigt, unter welchen Voraussetzungen die Anforderungen haftungsrechtliche Konsequenzen haben.

Material

Orientierungshilfe

Gemeinsame Verantwortlichkeit": Diese Orientierungshilfe des bayerischen Landesbeauftragten für den Datenschutz bietet eine Übersicht bezüglich der Voraussetzungen der gemeinsamen Verantwortlichkeit, Beispiele für gemeinsame Verantwortlichkeit, sowie Abgrenzungshinweise zu anderen Verarbeiterrollen.

Checkliste

Privacy by Design": Diese Checkliste soll einen Überblick über die in Art. 25 DSGVO statuierten Vorgaben an „Datenschutz durch Technikgestaltung und durch datenschutzfreundliche Voreinstellungen" bieten. Die Fragen werden ergänzt durch Informationen basierend Rechtsprechung, rechtswissenschaftlicher Literatur und Stellungnahmen deutscher und europäischer Aufsichtsbehörden.

Whitepaper

„Anforderungen für Ethikanträge" von Forschungsvorhaben„: Dieses Whitepaper gibt einen strukturierten Überblick über die rechtlichen und prozeduralen Anforderungen an die Einreichung von Ethikanträgen gemäß § 15 der Berufsordnung für die Ärztinnen und Ärzte des Saarlandes (BO-Ä). Es richtet sich an Forschungsvorhaben und erläutert die Zuständigkeiten, Antragsverfahren und einzureichenden Unterlagen.