

Transparenz, Erklärbarkeit und Vertrauen in KI-Systeme

4 STUNDEN

Einführung · Ablaufplan · Methoden · Material

Konzeption und Realisierung: K8 Institut für strategische Ästhetik gGmbH

Einführung & Relevanz

Transparenz, Erklärbarkeit und Vertrauen in KI-Systeme



Ich möchte verstehen, wie wir KI-Ergebnisse so transparent machen, dass Ärzt:innen, Patient:innen und Forschende ihnen vertrauen können.

Workshopinhalte

- Explainable AI (XAI) und Transparenzanforderungen in der Forschung
- Nachvollziehbarkeit und Dokumentation von Modellentscheidungen (Entscheidungsbäume und weitere?)
- Kommunikation von Unsicherheiten und Grenzen von KI-Systemen
- Aufbau von Vertrauen zwischen Technik, Medizin und Öffentlichkeit
- Praktische Methoden zur Bewertung der Erklärbarkeit

Learning Outcome

Die Teilnehmenden können beurteilen, wie erklärbar ein KI-System ist. Sie wissen, wie Vertrauen in KI-Forschung durch klare Kommunikation und Nachvollziehbarkeit gestärkt werden kann.

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 vorstellen; Spielregel: Der Workshop ersetzt keine technische Validierung einzelner Modelle	kein Material für diesen Baustein
	09:05–09:30 (25 Min.)	1-2-4-All (Liberating Structures) Methodenbeschreibung	– Leitfrage: „Wann hast du zuletzt einer KI-gestützten Diagnose oder Empfehlung vertraut oder misstraut – und warum?“ – Einzeln → Zweier → Vierer → Plenum	kein Material nötig – aktiviert Vorerfahrung
2. Vertiefung / intensive thematische Auseinandersetzung (100 Min.)	09:30–09:50 (20 Min.)	Kurzimpuls „Erklärbare KI (XAI) & kalibriertes Vertrauen“ (Plenum)	– Kalibriertes Vertrauen: weder Automation Bias (blindes Vertrauen) noch übertriebene Algorithm Aversion (Schlicker, Langer & Hirsch 2024) – Nachvollziehbarkeit von Modellentscheidungen: von Natur aus interpretierbare Modelle (u. a. Entscheidungsbäume) vs. Post-hoc-Erklärverfahren für Black-Box-Modelle (u. a. SHAP, LIME, kontrafaktische Erklärungen)	Schlicker, Langer & Hirsch (2024): Wie vertrauenswürdig ist künstliche Intelligenz?
	09:50–11:10 (80 Min.)	Critical Uncertainties (Liberating Structures) Methodenbeschreibung	2×2-Raster aus den Achsen „technische Erklärbarkeit“ (White-Box ↔️ Black-Box) und „kalibriertes Nutzer:innen-Vertrauen“ (hoch ↔️ niedrig); je Szenario Strategien für Dokumentation & Kommunikation entwickeln. – Besonders im Blick: das Szenario „Black-Box + geringes kalibriertes Vertrauen“ – hier besteht der größte Dokumentations- und Erklärungsbedarf Leitfrage: Wer muss was gegenüber wem dokumentieren/erklären – insbesondere bei institutionsübergreifenden Kooperationen?	Orientierungshilfe „Gemeinsame Verantwortlichkeit“ Whitepaper „Management von Betroffenenrechten“
Pause	11:10–11:25 (15 Min.)	Pause	—	—

Konzeption und Realisierung: K8 Institut für strategische Ästhetik gGmbH

Phase	Zeit	Methode / Format	Inhalt & Ziel	Material / Quelle
3. Diskussion / Austausch / Diskurs – Arbeiten mit dem Erfahrenen (60 Min.)	11:25–11:30 (5 Min.)	Rahmung durch Expert:in (Plenum)	– Kurze Vorstellung der/des Erfahrenen (Klinik-/Technikpraxis oder Patientenvertretung)	kein Material für diesen Baustein; ggfs. eigene Präsentation
	11:30–12:25 (55 Min.)	Wise Crowds (Liberating Structures) Methodenbeschreibung	– Eine Person schildert eine reale Vertrauens-/Erklärbarkeitsherausforderung; die ganze Gruppe berät gemeinsam – Klärungsfragen, dann Empfehlungen, während die Person zuhört – Bezug: Wie werden Erklärungs- und Verantwortungspflichten zwischen Projektpartnern im Kooperationsvertrag rechtssicher und pragmatisch geregelt?	Mustervereinbarungen für Forschungs- und Entwicklungskooperationen
4. Sicherung / Lernergebnis = Ausklang (35 Min.)	12:25–13:00 (35 Min.)	Empathy Map (Design Thinking, adaptiert) Methodenbeschreibung	– Kleingruppen erstellen je eine Empathy Map (Denken / Fühlen / Sagen / Tun) für eine Perspektive – Patient:in, Ärzt:in oder Entwickler:in – zum Thema Vertrauen in KI-Systeme – Plenum: Wo unterscheiden sich die Perspektiven – was bedeutet das für den Vertrauensaufbau zwischen Technik, Medizin und Öffentlichkeit? – Abschluss: persönliches Mitnahme-Learning, Dank & Hinweis auf alle Materialien als Handout	kein zusätzliches Material – thematischer Bezug zu Schlicker et al. (2024)

Methoden

Methoden & Formate

4 Formate für einen abwechslungsreichen und praxisorientierten Workshop.

1-2-4-All

Die Teilnehmenden reflektieren zunächst individuell, dann zu zweit, zu viert und schließlich in der gesamten Gruppe, welche Anforderungen sie an transparente und nachvollziehbare KI-Systeme stellen.

Critical Uncertainties

Zentrale Unsicherheiten beim Einsatz von KI-Systemen werden identifiziert und anhand möglicher Zukunftsszenarien untersucht. So lassen sich Herausforderungen für Transparenz, Erklärbarkeit und Vertrauen sichtbar machen.

Wise Crowds

Kollegiale Beratung zu konkreten Herausforderungen der KI-Entwicklung. Die Gruppe entwickelt gemeinsam Empfehlungen, wie Erklärbarkeit und transparente Kommunikation verbessert werden können.

Empathy Map

Perspektiven unterschiedlicher Zielgruppen werden betrachtet, beispielsweise von Ärzt:innen, Patient:innen und Forschenden. Im Workshop wird die Methode als Abschlussreflexion eingesetzt, um Anforderungen an verständliche KI-Kommunikation zusammenzuführen.

Material

Quellen & Arbeitsmaterialien

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

Schlicker, N., Langer, M. & Hirsch, M. C. (2024). Wie vertrauenswürdig ist künstliche Intelligenz? Forum, 39, 277–284.

Mustervereinbarungen für Forschungs- und Entwicklungskooperationen

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

Whitepaper „Management von Betroffenenrechten“

Material

Weiterführende Materialbeschreibungen

Erläuterungen zu den ergänzenden Quellen und Arbeitsmaterialien.

Mustervereinbarungen für Forschungs- und Entwicklungskooperationen“

Dies er Musterkonsortialvertrag zu Forschungskooperationen soll dabei als rechtlich geprüfte Leitlinie dienen und soweit möglich sicherstellen, dass bei der Zusammenarbeit der administrative Aufwand auf ein Mindestmaß reduziert wird, damit die Vereinbarung auch für kleinere und mittlere Hochschulen und Forschungseinrichtungen bzw. Unternehmen einfach verwendbar ist.

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.

Whitepaper

„Management von Betroffenenrechten“: Die Datenschutz-Grundverordnung (DSGVO) räumt betroffenen Personen eine Vielzahl von Rechten zum Schutz ihrer Daten ein. Im Rahmen ihrer Datenschutz-Compliance müssen Forschungsprojekte diese beachten. In diesem Whitepaper werden die einzelnen Betroffenenrechte erläutert und Tipps zur Erstellung und Einhaltung rechtskonformer Compliancestrukturen gegeben.