Vorschlag der EU-Kommission

AI HLEG: High Level Expert Group on AI
Ethische Richtlinien für vertrauenswürdige KI
ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence

I. Begründung in den Menschenrechten
II. Bedeutung für die Umsetzung
III. Überprüfbarkeit

Georg Ludwig Verhoeven
Integrata-Stiftung

Die AI HLEG, ihre Ziele und Dokumente

AI HLEG: High Level Group on Artifical Intelligence
52 Wissenschaftler*innen aus vielen EU-Staaten

Maximierung des Nutzens von AI bei gleichzeitiger Minimierung der Risiken

Ziel „Trustworthy AI made in Europe“ als Marke (anders als in USA und China), Europa als global führender AI-Innovator

Dokument 1: „Ethical Guidelines for Trustworthy AI“ - in Arbeit, Entwurf liegt vor (Dezember 2018), endgültige Fassung März 2019

Dokument 2: „AI Policy and Investment Recommendations“ – in Arbeit, geplant für Mai 2019

Diese Dokumente stellen keine Gesetze oder Verordnungen dar, sondern haben freiwliggen Charakter, zu dem sich die Organisationen verpflcihetn können / sollen.

Trustworthy AI – vertrauenswürdige AI

„human centric“, d. h. Entworfen, entwickelt und eingesetzt mit den Entwicklungsprinzip eines „ethical purpose“, basierend auf

- **Grundrechten:** Menschenwürde, Freiheit des Einzelnen
- **gesellschaftlichen Werten:** Demokratie, Gleichheit, keine Diskriminierung, Schutz von Minderheiten, Bürgerrechte

wie sie in den Statuten der EU festgelegt sind, dazu werden besonders berücksichtigt:

- Schutz von Schutzbedürftigen
- Situationen, bei denen eine asymmetrische Verteilung von Informationen / Daten vorliegt: Das AI-System „weiß“ mehr über die Objekte (Menschen) als die Objekte über das System

Wie erstellt man „Trustworthy AI“ ?

Ab der frühesten Phase des Systemspezifikation und des Designs sind – neben den fachlichen Anforderungen – zu berücksichtigen:

- klare Verantwortung der Ersteller / Nutzer
- verantwortlicher Umgang mit Daten
- Menschen als höchste Kontrollinstanz
- Vermeidung von Diskriminierung
- Schutz der Privatsphäre
- Technische Robustheit
- Sicherheit
- Transparenz / Nachvollziehbarkeit

- offene Kommunikation mit allen Beteiligten, Grenzen des Systems und realistische Erwartungen
- AI-Systeme als Teil der Unternehmenskultur, Verhaltensregeln für den Einsatz
- angemessene Ausbildung für alle beteiligten Gruppen (Management, Systemnutzer*innen, andere Mitarbeiter*innen)

- Beurteilungskriterien für AI-Systeme („Assessment“)

Technische Methoden zur Erstellung von Trustworthy AI (1)

Ethics & Rule of Law by Design: Bei jedem Designschritt wird neben der Funktionalität auch die Einhaltung von ethischen Regeln und Gesetzen berücksichtigt. Hiezu gehört auch „Privacy by Design“ und Security by Design“, insb. Schutz der Privatsphäre, Verhalten des Systems bei Ausfall und Wiederanlauf, Angriff von außen

Architektur: Die Architektur des Systems muss so angelegt sein, dass die Abläufe jederzeit den Grundregeln entsprechen und dass das System durch geeignete – ggf. externe – Komponenten überwacht und kontrolliert werden kann.

Test und Validierung: Tests müssen die Veränderung aller Komponenten – incl. Daten und externer Schnittstellen permanent kontrollieren; also permanentes und „adverses“ („feindliches“) Testen

Technische Methoden zur Erstellung von Trustworthy AI (2)

Traceability und Auditability: Entscheidungen des Systems müssen nachvollziehbar sein, das System muss die Schritte zu und Begründung für die Entscheidung dokumentieren (das ist heute technisch möglich, u. U. aufwändig)

Explanability (XAI: Explainable AI): In lernenden Systemen (> neuronale Netze) müssen die Lernprozesse (Justierung der Gewichte) nachvollziehbar sein (das ist noch Gegenstand der Forschung)

Nicht-technische Methoden zur Erstellung von Trustworthy AI

- Gesetze, Regeln und Vorschriften
- Normen und Standardisierung
- Führung, Management, Verantwortlichkeit, „Ethics Panel“
- Werte und Verhaltensregeln
- Ausbildung zu allen Aspekten des Einsatzes
- Offene Kommunikation mit allen Beteiligten
- Gemischte Design-Teams: Geschlecht, Kultur, Alter, berufliche Hintergünde und Kenntnisse

Beurteilung (Assessment) von Trustworthy AI

Wichtige Aspekte einer Bewertung für AI-Systeme auf der Basis der oben genannten Designanforderungen – diese Bewertung muss auf Basis gemachter Erfahrung ständig fortgeschrieben werden:

- **Verantwortung:** wer hat die Verantwortung, was passiert bei Probleme / Fehlern, Einflussmöglichkeiten von außen?
- **Daten- und Prozesskontrolle**
- **Design for all:** Ist das System für alle Nutzer*innen geeignet (insb. auch für „benachteiligte“ Personen, …
- **Kontrolle der Autonomität des Systems:** Kann das System jederzeit von außen kontrolliert werden, können Menschen die Kontrolle übernehmen?
- **Robustheit, Sicherheit:** gegenüber Angriffen von außen, Zuverlässigkeit, Reproduzierbarkeit, Fallback-Plan
- **Transparenz:** ist das Systemn durchsichtig / verstehbar – und insb. keine „black box“

Danke für Ihre Aufmerksamkeit!

Georg.Verhoeven@integrata-stiftung.de
www.humanithesia.org

integrata stiftung

Machen Sie mit!

Wecken Sie andere auf!
Diskutieren Sie mit Ihren Bekannten!

Kommen Sie zum Stuttgarter Zukunftssymposium:
www.stuttgarter-zukunftssymposium.de 15./16.11.2019

Oder melden Sie sich bei uns dafür:
info@integrata-stiftung.de
www.humnithesia.org