Judith Simon

Vertrauenswürdige KI? Zum Verhältnis von Vertrauen, Wissen und Technologie

20 Jahre TELOTA – Festveranstaltung am 22.06.2022

Berlin: Berlin-Brandenburgische Akademie der Wissenschaften, 2022

Persistent Identifier: urn:nbn:de:kobv:b4-opus4-38348

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Prof. Dr. Judith Simon

Vertrauenswürdige KI? Zum Verhältnis von Vertrauen, Wissen und Technologie

22.06.2022| 20 Jahre TELOTA – Festveranstaltung | BBAW | Berlin

Aufbau

1. Vertrauen in KI - Vertrauenswürdige KI?
2. Vertrauen & Vertrauenswürdigkeit: Philosophische Perspektiven
   - Ethik: Annette Baier (1985): Trust and Antitrust
   - Wissenschaftstheorie: John Hardwig (1991): The Role of Trust in Knowledge
3. Fazit: Vertrauen in KI - Vertrauenswürdige KI?

- Rieder, G., Simon, J. Wong P-H. (2021). Mapping the Stony Road toward Trustworthy AI: Expectations, Problems, Conundrums. Machines We Trust: Perspectives on Dependable AI. M. Pelillo and T. Scantamburlo. Cambridge, MA, MIT Press. DOI: http://dx.doi.org/10.2139/ssrn.3717451 & https://ssrn.com/abstract=3717451

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Vertrauen in KI – Vertrauenswürdige KI?

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Vertrauenswürdige **KI**

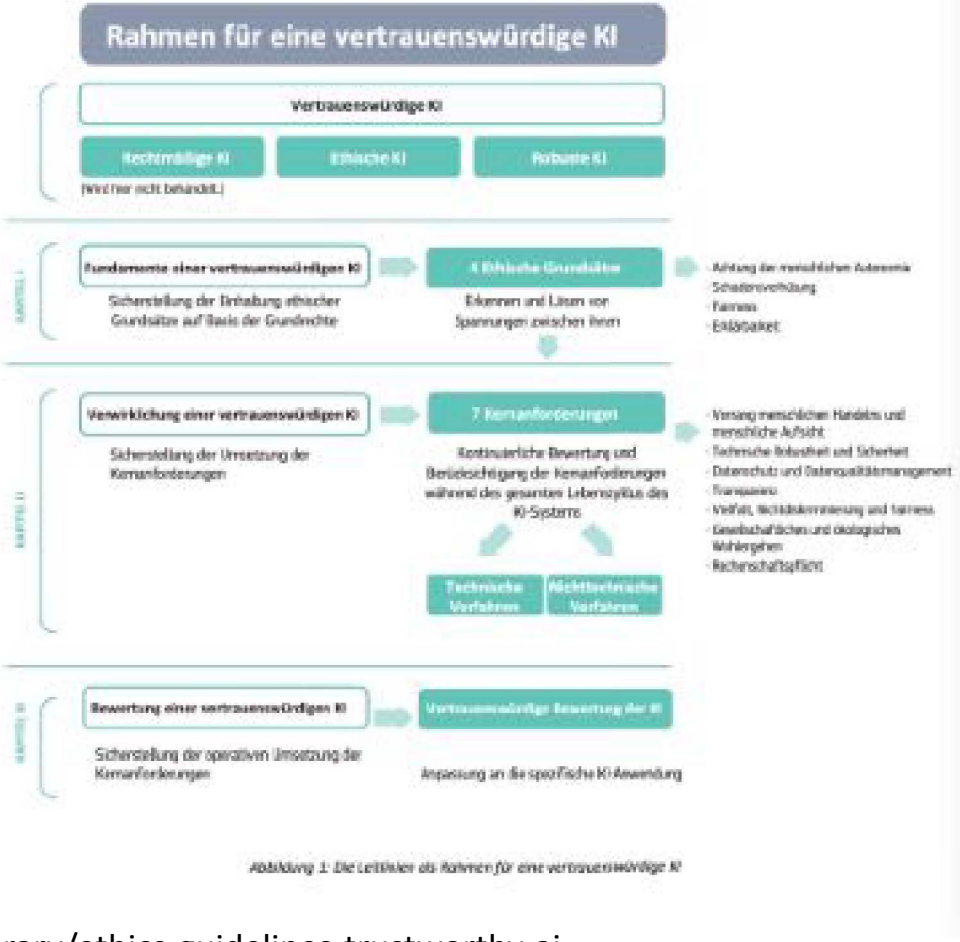


Abbildung 1: Die Leitlinien als Rahmen für eine vertrauenswürdige KI

https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Künstliche Intelligenz

"Artificial Intelligence [...] is the subfield of Computer Science devoted to developing programs that enable computers to display behavior that can (broadly) be characterized as intelligent. Most research in AI is devoted to fairly narrow applications, such as planning or speech-to-speech translation in limited, well defined task domains. But substantial interest remains in the long-range goal of building generally intelligent, autonomous agents, even if the goal of fully human-like intelligence is elusive and is seldom pursued explicitly and as such."

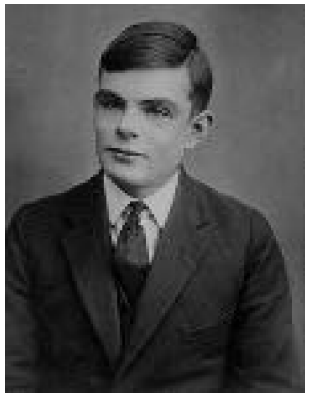
(Thomason 2013: https://plato.stanford.edu/entries/logic-ai/)

Künstliche Intelligenz

- Starke KI/Breite/Generelle KI (Strong AI, AGI)
  - Maschinen deren allgemeine Intelligenz sich nicht von menschlicher Intelligenz unterscheidet oder diese übertrifft
  - Maschinen, die alle Aspekte menschlicher KI perfekt simulieren können
- **Schwache KI (Weak AI)**
  - **Technische Lösung spezieller kognitiver Teilprobleme: Spracherkennung, Bilderkennen/Computer Vision, Robotik, Expertensysteme,...**

Künstliche Intelligenz: lange Geschichte mit Höhen und Tiefen

Bernhard Christoph Francke



Turing/Public Domain

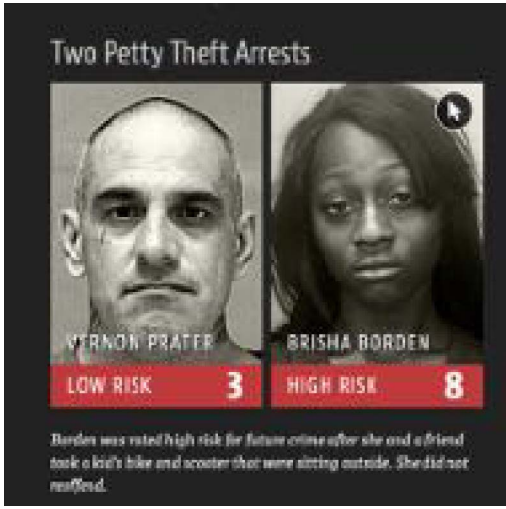
Olivier Carré-Delisle/Flickr

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Daten, Künstliche Intelligenz & Ethik

- Kern: Analyse großer Datenmengen > Mustererkennung, Klassifikation, Prognose & Entscheidungsvorbereitung
- Mannigfaltigkeit, Komplexität & Dynamik
  1. der Technologien
  2. der Entwicklungs- und Anwendungskontexte
  3. der beteiligten Akteure
  4. der ethischen Fragen
  5. der notwendigen Regulierung

  → Ökosystemperspektive auf Daten & KI



1) Apple, Facebook Google, Amazon
2) Propublica
3) https://www.mdr.de/wissen/mensch-alltag/die-verschiedenen-plaene-der-laender-fuer-eine-corona-app-100.html

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Vertrauenswürdige KI

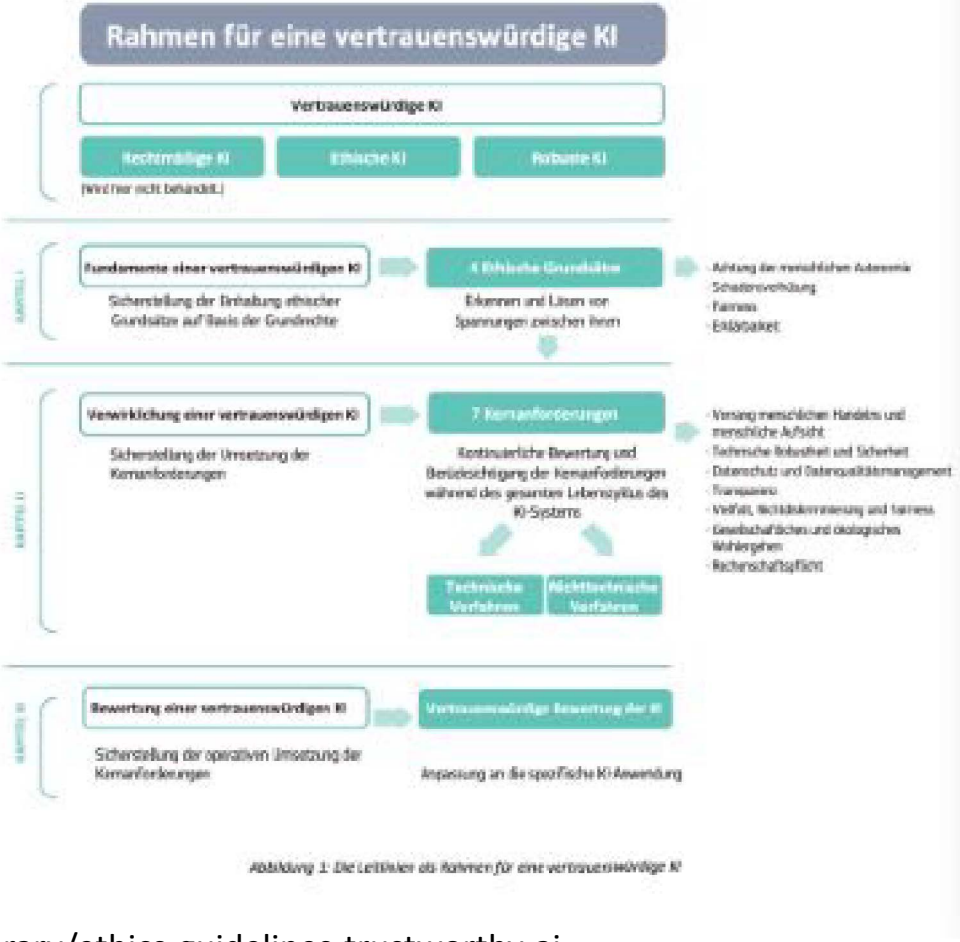


Abbildung 1: Die Leitlinien als Rahmen für eine vertrauenswürdige KI

https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

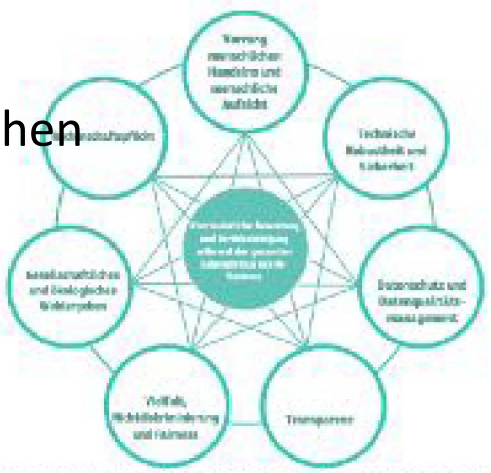
- 4 Ethische Grundsätze im Kontext von KI-Systemen
  1. Achtung der menschlichen Autonomie
  2. Schadensverhütung
  3. Fairness
  4. **Erklärbarkeit**

https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Erklärbarkeit ist unabdingbar, wenn beim Benutzer dauerhaftes Vertrauen in KI-Systeme entstehen soll. Das bedeutet, dass Prozesse transparent sein müssen, dass die Fähigkeiten und der Zweck von KI-Systemen offen zu kommunizieren sind und dass Entscheidungen – im größtmöglichen Umfang – den direkt und indirekt davon betroffenen Personen erklärbar sein müssen. Ohne diese Informationen kann eine Entscheidung nicht ordnungsgemäß angefochten werden. Eine Erklärung, warum ein Modell ein bestimmtes Ergebnis oder eine bestimmte Entscheidung erzeugt hat (und welche Kombination aus Eingabefaktoren dazu geführt hat) ist nicht immer möglich. Diese Fälle werden als „Blackbox"-Algorithmen bezeichnet und erfordern besondere Beachtung. Unter diesen Umständen sind möglicherweise andere Erklärbarkeitsmaßnahmen notwendig (z. B. Rückverfolgbarkeit, Nachprüfbarkeit und transparente Kommunikation über die Fähigkeiten des Systems), solange das System als Ganzes Grundrechte achtet. Bis zu welchem Grad Erklärbarkeit notwendig ist, hängt sehr stark vom Kontext und der Tragweite der Konsequenzen eines fehlerhaften oder anderweitig unzutreffenden Ergebnisses ab.

https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

- 7 Anforderungen an vertrauenswürdige KI
  1. Vorrang menschlichen Handelns und menschlicher Aufsicht
  2. Technische Robustheit und Sicherheit
  3. Schutz der Privatsphäre und Datenqualitätsmanagement
  4. Transparenz
     - Z.B. Nachverfolgbarkeit, Erklärbarkeit und Kommunikation
  5. Vielfalt, Nichtdiskriminierung und Fairness
  6. Gesellschaftliches und ökologisches Wohlergehen
  7. Rechenschaftspflicht



https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

- Rückverfolgbarkeit
  - Dokumentation von Datensätzen und Prozessen, inkl. Datenerhebung, Labeling, verwendete Algorithmen, ...
- Erklärbarkeit
  - Möglichkeit die technischen Prozesse eines KI Systems und die damit verbundenen menschlichen Entscheidungen zu erklären
  - Erfordert dass Entscheidungen eines KI Sytems von Menschen verstanden und rückverfolgt werden können
  - Ggf. Trade-offs nötig. Z.B. zwischen Genauigkeit und Erklärbarkeit
- Kommunication
  - Recht zu wissen, ob man mit KI interagiert
  - Kennzeichnung der Fähigkeiten und Grenzen von KI Systemen

http s://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Vertrauen und Vertrauenswürdigkeit

- Zwei Fragen zu Vertrauenswürdiger KI/Vertrauen in KI
  1. Können wir KI vertrauen?
     - Was ist Vertrauen? Wer kann wem vertrauen?
  2. (Wann) sollten wir KI vertrauen?
     - Beziehungen zwischen Vertrauen, Vertrauenswürdigkeit, Erklärungen und Wissen

- Einsichten aus **Ethik** und Wissenschaftstheorie

"Whatever matters to human beings, trust is the atmosphere in which it thrives."

Bok (1978), in Baier (1986)

" Exploitation and conspiracy, as much as justice and fellowship, thrive better in an atmosphere of trust.

"Trust then […] is accepted vulnerability to another's possible but not expected ill will (or lack of good will) toward one."

"Trust is a fragile plant, which may not endure inspection of its roots, even when they were, before the inspection, quite healthy."

Baier (1986) : Trust and Antitrust

- Vertrauen
  - als akzeptierte Vulnerabilität
    - Vertrauen <> Sicherheit, Gewissheit
  - als relationales Konzept: A vertraut B in Bezug auf X
    - Ich vertraue meiner Ärztin in Bezug auf medizinische Diagnose, nicht Reparatur meines Autos...
  - beschreibt sehr verschiedene Relationen in
    - Personen (Partner, Kind, Fremde, Hausärztin..)
    - Institutionen (Bundesregierung, STIKO, ÖRR ...)
    - abstrakte Entitäten (die Wissenschaft, die Politik, die Medien, ..)
    - Technologien? Künstliche Intelligenz?
    - ...
  - wird (erst) zum Thema, wenn es in Frage gestellt wird
  - ist leicht verloren und schwer wieder herzustellen

- Zwei Fragen zu Vertrauenswürdiger KI/Vertrauen in KI
  1. Können wir KI vertrauen?
     - Was ist Vertrauen? Wer kann wem vertrauen?
  2. (Wann) sollten wir KI vertrauen?
     - Beziehungen zwischen Vertrauen, Vertrauenswürdigkeit, Erklärungen und Wissen

- Einsichten aus Ethik und **Wissenschaftstheorie**

- Vertrauen und Wissen/schaft

"It seems paradoxical that scientific research, in many ways one of the most questioning and skeptical of human activities, should be dependent on personal trust. But the fact is that without trust the research enterprise could not function.. . .Research is a collegial activity that requires its practitioners to trust the integrity of their colleagues."

Relman in Hardwig (1991)

Hardwig (1991). The Role of Trust in Knowledge

- Vertrauen und Wissen/schaft

- Analyse von wissenschaftlicher Erkenntnisproduktion in Mathematik & Physik
- Arbeitsteilung aufgrund unterschiedlicher Expertise und Zeitersparnis

→ Wer weiß etwas? Wer ist Träger wissenschaftlichen Wissens?

Hardwig (1991). The Role of Trust in Knowledge

- Vertrauen und Wissen/schaft
  - Ist Wissen/schaft nicht das Gegenteil von Vertrauen?
  - Erkenntnistheorie: Ideal des selbstständigen Erkennens, Kritik an blindem Vertrauen in Autoritäten, in das Zeugnis anderer

- Vertrauen und Wissen/schaft

"Modern knowers cannot be independent and self-reliant, not even in their own fields of specialization. In most disciplines, those who do not trust cannot know; those who do not trust cannot have the best evidence for their beliefs. In an important sense, then, trust is often epistemologically even more basic than empirical data or logical arguments: the data and the argument are available only through trust. If the metaphor of foundation is still useful, the trustworthiness of members of epistemic communities is the ultimate foundation for much of our knowledge."

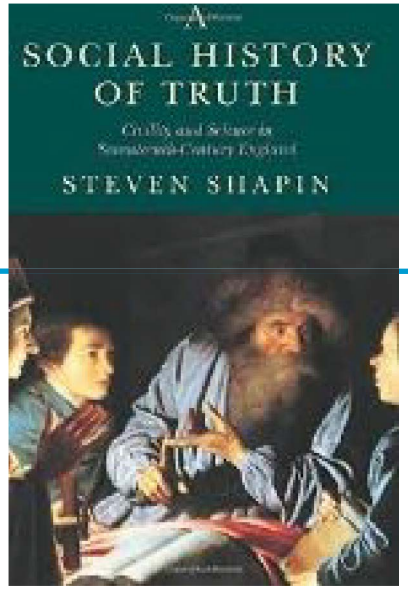
Hardwig (1991). The Role of Trust in Knowledge

- Vertrauen und Wissen/schaft

- Wissenschaftler müssen sich gegenseitig vertrauen in Bezug auf ihre
  - Kompetenz
  - Ehrlichkeit
  - Angemessene epistemische Selbsteinschätzung

- Epistemische und moralische Komponenten von Vertrauen und Vertrauenswürdigkeit

Hardwig (1991). The Role of Trust in Knowledge



- Vertrauen und Wissen/schaft

"[k]nowledge is a collective good. In securing our knowledge we rely upon others, and we cannot dispense with that reliance. That means that the relations in which we have and hold our knowledge have a moral character, and the word I use to indicate that moral relation is trust" (Shapin: XXV)

Shapin (1994): „A social history of truth"

- Vertrauen und Wissen/schaft

- Vertrauenswürdig = unabhängig + frei = Gentlemen
- Gentlemen waren finanziell unabhängig und frei, sie mussten nicht arbeiten
- Ökonomische Freiheit als Voraussetzung für moralische Freiheit
- Sozialer Status → ethischer Status → epistemischer Status von Aussagen
- Frauen wurden nicht als vertrauenswürdig angesehen, da sie ökonomisch abhängig waren

Shapin (1994): „A social history of truth“

- Vertrauen und Wissen/schaft

- Gegenthese: Gentlemen vertrauten sich nicht (nur) aufgrund ihrer (vorgeblichen) ökonomischen Unabhängigkeit, Tugend oder Ehre, sondern, aufgrund der engen sozialen Bezüge, d.h. weil sie sich gut kannten:

„Faced then with the question of why gentlemen trusted each other, the obvious answer is simply that they all belonged to the same club."

Lipton (1998): „The epistemology of testimony"

(Feministische) Kritik

- Wenn soziale Kriterien verwendet werden, um epistemische Vertrauenswürdigkeit von Akteuren abzuschätzen, öffnet dies Tür & Tor für soziale Ungerechtigkeit und Diskriminierung

“testimonial injustice occurs when prejudice causes a hearer to give a deflated level of credibility to a speaker’s word” (Fricker 2007: 1).”

Fricker (2007) „Epistemic Injustice: Power and the Ethics of Knowing“

- Ist Vertrauen dann immer gut?
  - Nein, denn Vertrauen ist fehlbar und wir machen uns durch fehlgesetztes Vertrauen verwundbar

- 2 Fehler des Vertrauens
  - α-Fehler des Vertrauens: man vertraut denen, die nicht vertrauenswürdig sind
  - β-Fehler des Vertrauens: man vertraut nicht, obwohl jemand/etwas vertrauenswürdig gewesen wäre

  → Epistemische, moralische und praktische Schäden

- Wann/wem sollten wir vertrauen?
  - Vertrauen denen und nur denen, die vertrauenswürdig sind
- Vertrauenswürdige Akteure sind
  - kompetent, ehrlich, kennen die Grenzen ihrer Kompetenz (Hardwig 1991)
  - dem Vertrauenden gegenüber wohlwollend (Baier 1986)
- Wie können wir selbst vertrauenswürdig sein?
  - Epistemisch: Kompetent sein, aber auch die Grenzen der eigenen Kompetenz erkennen und signalisieren
  - Moralisch: ehrlich und wohlwollend sein

Fazit

- Vertrauen in nicht prinzipiell gut, sondern nur wenn es in vertrauenswürdige Akteure/Systeme gesetzt wird
- Doppelte Gefahr: α-Fehler & β-Fehler des Vertrauens

- Vertrauen in KI = Vertrauen in KI als sozio-technischem System
- Damit das Vertrauen in KI gerechtfertigt sein kann, müssen KI/ADM-Systeme vertrauenswürdig sein
- KI/ADM Systeme können nur vertrauenwürdig sein als sozio-technische Ökosysteme

- Vertrauenswürdigkeit hat epistemische und moralische Dimensionen 1. und 2. Ordnung
- Epistemisch:
  - Epistemisch[1]: kompetent sein
  - Für KI: hohe Genauigkeit und Robustheit, d.h. das gut und zuverlässig messen, vorhersagen, was man vorhersagen will, Biases minimieren, ...
  - Epistemisch[2]: die Grenzen der eigenen Expertise kennen und offenlegen
  - Für KI: Fehlerrate, Biases, Limitationen deutlich erkennbar machen

- Vertrauenswürdigkeit hat epistemische und moralische Dimensionen 1. und 2. Ordnung
- Moralisch
  - Moralisch[1]: Ehrlich, guten Willens sein, ...
  - Für KI: Guter Zweck, ...
  - Moralisch[2]: Wissen, was eine moralische und keine technische, Frage ist und daher ggf. einer breiteren Beteiligung/weiterer Expertise bedarf
  - Für KI: Welches sind angemessene Methoden um Verzerrungen aufzudecken und zu mitigieren, angemessene fairness metrics, ...

- Vertrauenswürdige Erklärungen sind notwendig für gerechtfertigtes Vertrauen in KI/ADM Systeme
- Die Anforderungen an Erklärbarkeit können je nach System und Anwendungskontext variieren
  - Zwischen 99,9% Genauigkeit & 99,9% Erklärbarkeit
- Vertrauenswürdige Erklärungen müssen a) genau und b) angemessen sein, d.h. den Kontext,- Aufgaben und –Adressatenspezifischen Anforderungen genügen
  - Kontext: e.g. Sektor, hohe/niedriger Impact, …
  - Aufgabe: Auskunftspflicht nach DSGVO, Auditierung, Debugging/Security Check, ...
  - Adressat: Kunde/Patient, Arzt, Entwickler, Aufsichtsbehörde, NGO, ...

- Rieder, G., Simon, J. Wong P-H. (2021). Mapping the Stony Road toward Trustworthy AI: Expectations, Problems, Conundrums. Machines We Trust: Perspectives on Dependable AI. M. Pelillo and T. Scantamburlo. Cambridge, MA, MIT Press. DOI: http://dx.doi.org/10.2139/ssrn.3717451 & https://ssrn.com/abstract=3717451

- Asghari, H. B., N.; Burchardt, A.; Dicks, D.; Faßbender, J.; Feldhus, N.; Hewett, F.; Hofmann, V.; Kettemann, M. C.; Schulz, W.; Simon, J.; Stolberg-Larsen, J.; and Züger, T. (2021). "What to explain when explaining is difficult? An interdisciplinary primer on XAI and meaningful information in automated decision-making." Alexander von Humboldt Institute for Internet and Society. DOI: https://doi.org/10.5281/zenodo.6375784.

- Simon, J., Ed. (2020). The Routledge Handbook of Trust and Philosophy. The Routledge Handbooks in Philosophy. New York, Taylor & Francis.

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Vielen Dank für Ihre Aufmerksamkeit!

Prof. Dr. Judith Simon
Professorin für Ethik in der Informationstechnologie
Universität Hamburg
Email: judith.simon@uni-hamburg.de
Web: https://www.inf.uni-hamburg.de/en/inst/ab/eit/team/simon.html