Digitale Souveränität und Künstliche Intelligenz für den Menschen

Roland Vogt(✉)
Labor für Zertifizierung und Digitale Souveränität, Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI), Saarland Informatics Campus D3 2, 66123 Saarbrücken, Deutschland
roland.vogt@dfki.de

Zusammenfassung. Die Förderung der digitalen Souveränität von Personen, die von der Verwendung von Systemen der Künstliche Intelligenz (KI) betroffen sind, wird nur gelingen, wenn die KI selbst die Handlungsfähigkeit und -kompetenz des einzelnen Menschen unterstützt, statt sie seiner Kontrolle zu entziehen. Dies verlangt die Gestaltung von KI für den Menschen mit eingebauter digitaler Souveränität (Sovereignty by Design). Dafür spielt die Schaffung von Rahmenbedingungen aus Standardisierung und Bewertung eine zentrale Rolle.

Schlüsselwörter: Digitale Souveränität · Künstliche Intelligenz · Sovereignty by Design

1 Einleitung

Wenn eine neue Technologie boomt, entstehen innovative Produkte, Dienstleistungen, Prozesse, Strukturen und Kompetenzen. Diese Innovationen führen zu erheblichen Veränderungen in ihren Anwendungsbereichen, verbunden mit der Verbesserung, Umgestaltung oder Ablösung von bisher bewährten Technologien. Die Veränderungen reichen umso weiter, je breiter ihre Anwendungsbereiche gestreut sind.

Eine solche Technologie ist die Künstliche Intelligenz (KI). Angetrieben durch zunächst vereinzelte herausragende Erfolgsgeschichten im maschinellen Lernen als wichtigem Teilgebiet der KI, wie etwa für komplexe Strategiespiele oder autonome Fahrzeuge, werden derzeit durch die breite Verwendung von KI systemische Veränderungen in den unterschiedlichsten Anwendungsbereichen eingeleitet. Die Anwendungsmöglichkeiten von KI werden überall da erschlossen, wo durch Digitalisierung große Mengen an Daten verarbeitet werden. Sie greift dabei signifikant in die grundlegenden Prinzipien globaler Infrastrukturen wie Verkehr, Energie, Industrie, Agrarwirtschaft, Kreditwirtschaft, Gesundheit und Verwaltung ein.

Begleitet wird der KI-Boom von einem breiten politischen und gesellschaftlichen Diskurs über die Chancen, aber auch die Risiken der beginnenden Veränderungen. Über die Glaubwürdigkeit und Vertrauenswürdigkeit von KI wird ebenso diskutiert wie über gesetzliche und ethische Beschränkungen ihres Einsatzes. Die Diskussion wird kontrovers und emotional geführt. Eine Versachlichung der Argumentation

E. A. Hartmann (Hrsg.): *Digitalisierung souverän gestalten,* S. 143–150, 2021.
https://doi.org/10.1007/978-3-662-62377-0_11

ist schwierig, weil das Verhalten von KI-Systemen häufig auch von Experten nur unzureichend verstanden und erklärt werden kann. So ist etwa nur bedingt vorhersehbar, ob ein autonomes Fahrzeug bestimmte Situationen falsch einschätzt und damit das Leben von Verkehrsbeteiligten gefährdet. Dies stellt auch Versicherungen vor neuartige Herausforderungen für die Kalkulation von Risiken.

Die wissenschaftliche Forschung stellt sich diesen Herausforderungen durch Intensivierung und Vernetzung der internationalen Zusammenarbeit. Erwähnt sei hier die Initiative Confederation of Laboratories for Artificial Intelligence Research in Europe (CLAIRE) [1], ein paneuropäisches Bündnis von Forschungslaboren für KI in Europa. CLAIRE wird sich auf eine vertrauenswürdige KI konzentrieren, die die menschliche Intelligenz fördert, anstatt sie zu ersetzen, und die somit jedem einzelnen Menschen zugutekommt.

Die Förderung von vertrauenswürdiger KI für den einzelnen Menschen ist eng verknüpft mit der digitalen Souveränität, das heißt, mit der Möglichkeit und Fähigkeit des einzelnen Menschen, KI-Technologie kompetent und zielgerichtet so einzusetzen, dass sie die eigene Handlungsfähigkeit und -kompetenz unterstützt, statt sie seiner Kontrolle zu entziehen.

2 KI für den Menschen

Ausgangspunkt unserer Betrachtung sind Richtlinien für ethische und vertrauenswürdige KI, die in großer Zahl von verschiedensten Organisationen veröffentlicht worden sind [2]. Ein vergleichender Überblick dieser Richtlinien ist kaum möglich. Und dies ist auch nicht unser Ziel. Wir konzentrieren uns hier vielmehr auf den Vorschlag einer von der Europäischen Kommission eingesetzten, unabhängigen, internationalen Expertengruppe aus unterschiedlichen Disziplinen. Die von dieser Expertengruppe verfassten Ethischen Richtlinien für vertrauenswürdige KI [3] zeichnen sich durch die fachliche Exzellenz der Expertengruppe einerseits und die Ausstrahlung auf die gesamte Europäische Union andererseits aus.

Die Richtlinien [3] der Europäischen Kommission skizzieren einen Rahmen für vertrauenswürdige KI und zielen dabei auf die Gewährleistung ethischer Grundsätze und Werte sowie auf die Gewährleistung der Robustheit aus einer technischen und gesellschaftlichen Perspektive. Sie beschäftigt sich ausdrücklich nicht mit rechtlichen Rahmenbedingungen für den Einsatz von KI-Technologie. Bei unserer Auswertung stehen die Auswirkungen der Verwendung von KI auf den Menschen im Fokus der Untersuchung, weshalb die Ausblendung rechtlicher Fragestellungen angemessen ist.

Aus der Charta der Grundrechte der Europäischen Union [4] werden in den Richtlinien [3] zusammenfassend die folgenden vier ethischen Grundsätze für vertrauenswürdige KI hergeleitet:

1. Achtung der Autonomie des Menschen (Selbstbestimmung)
2. Abwendung von Schaden (mentale und körperliche Unversehrtheit)
3. Fairness (Ausgleich der Interessen)
4. Erklärbarkeit (Nachvollziehbarkeit und Vorhersagbarkeit)

Durch die Orientierung an den Grundrechten der EU wird deutlich, dass die Grundsätze und Anforderungen der Richtlinien [3] den einzelnen Menschen ins Zentrum der Betrachtung stellen. KI für den Menschen kann damit als Leitgedanke der Richtlinien bezeichnet werden.

Aus diesen vier Grundsätzen werden sodann konkrete Anforderungen abgeleitet, wie in Tab. 1 dargestellt.

Tab. 1. Ethical requirements for trustworthy Artificial Intelligence [3]

Requirement	With elements
Human agency and oversight	Fundamental rights Human agency Human oversight
Technical robustness and safety	Resilience to attack and security Fall back plan and general safety Accuracy Reliability and reproducibility
Privacy and data governance	Respect for privacy and data protection Quality and integrity of data Access to data
Transparency	Traceability Explainability Communication
Diversity, non-discrimination and fairness	Avoidance of unfair bias Accessibility and universal design Stakeholder participation
Societal and environmental wellbeing	Sustainability and environmental friendliness Social impact Society and democracy
Accountability	Auditability Minimisation and reporting of negative impact Trade-offs Redress

3 Digitale Souveränität und KI-Systeme

Wir betrachten die im vorigen Abschnitt angeführten Grundsätze und Anforderungen im engeren Sinn als Maßstab für die Eigenschaften von KI-Systemen. Für ihre Ausgestaltung stellen wir sie in Zusammenhang mit der digitalen Souveränität des einzelnen Menschen und schlagen eine Kategorisierung vor, die bereits etablierte Konzepte einbindet und erweitert.

Zwischen den Grundsätzen für vertrauenswürdige KI und der digitalen Souveränität des einzelnen Menschen besteht ganz offensichtlich ein enger Zusammenhang. Die Achtung der Autonomie des Menschen adressiert unmittelbar die Erhaltung seiner Handlungsfähigkeit und -kompetenz. Dieses Kernelement der digitalen Souveränität wird flankiert von den weiteren Grundsätzen der Vertrauenswürdigkeit

(Unversehrtheit, Fairness und Erklärbarkeit), ohne deren Einhaltung ein souveräner Einsatz von KI-Technologie eingeschränkt oder verhindert wird.

Für den Einsatz von Informationstechnologie gibt es etablierte Konzepte für Sicherheit (sowohl security als auch safety) und informationelle Selbstbestimmung (privacy). Diese Konzepte können auch für KI-Technologien angewendet werden. Wir beschreiben zunächst die Kategorien *security* und *privacy,* und schlagen vor, *safety* in den umfassenderen Kontext der Kategorie *autonomy* zu stellen. Für jede dieser Kategorien stellen wir ihre Beziehung zu den im vorigen Abschnitt aufgeführten Grundsätzen und Anforderungen an die Vertrauenswürdigkeit von KI-Systemen dar.

3.1 Security

Für die Kategorie *security* gibt es schon seit den Anfängen der digitalisierten Datenverarbeitung allgemein anerkannte, übergeordnete Ziele, die hier kurz beschrieben werden.

1. Vertraulichkeit
   Daten werden in einer Weise verarbeitet, die einen dem Risiko angemessenen Schutz vor unbefugter oder unrechtmäßiger Verarbeitung gewährleistet.
2. Integrität
   Daten werden in einer Weise verarbeitet, die einen dem Risiko angemessenen Schutz vor unbeabsichtigtem Verlust, unbeabsichtigter Zerstörung oder unbeabsichtigter Schädigung gewährleistet.
3. Verfügbarkeit
   Daten werden in einer Weise verarbeitet, die gewährleistet, dass sie für die Verarbeitungsvorgänge auffindbar, darstellbar und interpretierbar sind.
4. Belastbarkeit
   Daten werden in einer Weise verarbeitet, die gewährleistet, dass sie bei einem materiellen oder technischen Zwischenfall rasch wiederhergestellt werden.

Die beschriebenen Ziele leisten einen wesentlichen Beitrag zu den Grundsätzen

- Achtung der Autonomie des Menschen: Vertraulichkeit, Integrität, Verfügbarkeit, Belastbarkeit,
- Abwendung von Schaden: Integrität, Verfügbarkeit, Belastbarkeit,
- Fairness: Vertraulichkeit, Integrität, Verfügbarkeit, Belastbarkeit.

3.2 Privacy

Für die Kategorie *privacy* gibt es aus der Perspektive des Schutzes der Verarbeitung personenbezogener Daten allgemein anerkannte, übergeordnete Ziele, die aus dem Standarddatenschutzmodell [5] übernommen und hier kurz beschrieben werden.

1. Transparenz
   Personenbezogene Daten werden in einer für die betroffenen Personen, für die verantwortlichen Beschäftigten, für die Datenschutzbeauftragten und für die Aufsichtsbehörden nachvollziehbaren Weise verarbeitet.

2. Interventionsbefähigung
   Personenbezogene Daten werden getreu der berechtigten Intervention der betroffenen Person berichtigt, vervollständigt, gelöscht, gesperrt oder übertragen, und sie werden nicht mehr verarbeitet, wenn die Rechtmäßigkeit der Verarbeitung durch Widerruf der Einwilligung oder berechtigten Widerspruch aufgehoben wird.
3. Zweckbindung
   Personenbezogene Daten werden für festgelegte, eindeutige und legitime Zwecke verarbeitet, einschließlich gegebenenfalls zulässiger Zweckänderungen, etwa zur Weiterverarbeitung für wissenschaftliche Forschungszwecke.
4. Datenminimierung
   Personenbezogene Daten werden nur verarbeitet, soweit und solange sie für die Zwecke angemessen, erheblich und auf das notwendige Maß beschränkt sind.

Die beschriebenen Ziele leisten einen wesentlichen Beitrag zu den Grundsätzen

- Achtung der Autonomie des Menschen: Transparenz, Interventionsbefähigung, Zweckbindung, Datenminimierung,
- Fairness: Transparenz, Interventionsbefähigung, Zweckbindung, Datenminimierung,
- Erklärbarkeit: Transparenz, Zweckbindung.

3.3 Autonomy

Obgleich die etablierten Ziele für die Kategorien *security* und *privacy* zur Umsetzung aller vier Grundsätze für vertrauenswürdige KI beitragen, decken sie doch nur bestimmte Aspekte der Datenverarbeitung ab. Die hier vorgeschlagenen Ziele für die Kategorie *autonomy* fokussieren direkt auf die Handlungsfähigkeit und -kompetenz und adressieren damit wesentliche Aspekte der digitalen Souveränität des einzelnen Menschen.

1. Kontrollbefähigung
   Das KI-System arbeitet in einer Weise, die den einzelnen Menschen dabei unterstützt, selbstbestimmte Entscheidungen zu treffen und in Übereinstimmung mit seinen Zielen zu handeln. Diese Fähigkeit begegnet unerwünschten Entscheidungen und fördert die Autonomie des Menschen.
2. Nachvollziehbarkeit
   Das KI-System arbeitet in einer Weise, dass seine Entscheidungen oder sein Verhalten für die von den Auswirkungen betroffenen Personen nachvollziehbar (erklärbar oder vorhersagbar) sind. Diese Fähigkeit begegnet fehlerhaftem oder unangemessenem Verhalten und unterstützt die Transparenz und Glaubwürdigkeit von KI-Technologie.
3. Beständigkeit
   Das KI-System arbeitet in einer Weise, dass ähnliche Aufgaben zu ähnlichen Ergebnissen (Entscheidungen oder Verhalten) führen. Diese Fähigkeit begegnet ungleichmäßigem, unerwartetem und unvorhersehbarem Verhalten und unterstützt die Verlässlichkeit und Glaubwürdigkeit von KI-Technologie.

4. Sicherheit (safety)
   Das KI-System arbeitet in einer Weise, die den einzelnen Menschen und seine Umwelt vor Schaden schützt. Diese Fähigkeit begegnet benachteiligendem oder verletzendem Verhalten und unterstützt die mentale und physische Unversehrtheit des Menschen.

Die beschriebenen Ziele leisten einen wesentlichen Beitrag zu den Grundsätzen

- Achtung der Autonomie des Menschen: Kontrollbefähigung,
- Abwendung von Schaden: Sicherheit (safety), Beständigkeit,
- Fairness: Kontrollbefähigung, Beständigkeit,
- Erklärbarkeit: Beständigkeit, Nachvollziehbarkeit.

4 Sovereignty by Design

Jede im vorigen Abschnitt beschriebene Kategorie, also *security, privacy* und *autonomy,* sollte von KI-Systemen aus sich selbst heraus angestrebt werden. Das Ziel besteht somit darin, KI-Systeme mit eingebauter digitaler Souveränität zu gestalten (Sovereignty by Design). Dazu werden zunächst technische Strategien, Konzepte und Muster für die Gestaltung benötigt. Diese sollten von Standardisierung begleitet werden, um eine vergleichende Einordnung und Bewertung, etwa durch Zertifizierung, zu ermöglichen.

4.1 KI-Systeme mit eingebauter digitaler Souveränität (Sovereignty by Design)

Für Security by Design existieren bewährte und praxiserprobte Konzepte für verschiedene Sicherheitsarchitekturen und Kategorien von IT-Systemen. Hier seien exemplarisch die Prinzipien des Open Web Application Security Project (OWASP) [6] für Security by Design von Webanwendungen genannt. Viele dieser Prinzipien lassen sich auf die Gestaltung von KI-Systemen anpassen.

Für Privacy by Design existieren Empfehlungen für Strategien, Konzept und Muster. Einen guten Überblick gibt ein Bericht der Agentur der Europäischen Union für Cybersicherheit ENISA [7] mit einer Bestandsaufnahme unter Berücksichtigung der rechtlichen Rahmenbedingungen. Auch diese Empfehlungen, insbesondere die darin enthaltenen prozess- und datenorientierten Strategien, lassen sich auf die Gestaltung von KI-Systemen anpassen.

Für Autonomy by Design können die hier vorgeschlagenen Kategorien einen Rahmen für die Entwicklung von Konzepten bilden. Vorhandene wissenschaftliche Ansätze, die etwa in den Richtlinien [3] der Europäischen Kommission beschrieben sind, können in diesen Rahmen eingebettet werden. Durch praktische Erprobung und wissenschaftliche Aufbereitung kann das Konzept reifen.

4.2 Standardisierung und Zertifizierung

Für die Kategorie *security* existieren eine Vielzahl von internationalen technischen Standards. Soweit sie querschnittliche Technologien, wie etwa kryptographische Algorithmen und Protokolle, adressieren, können sie uneingeschränkt für die Gestaltung von KI-Technologie verwendet werden. Auch für die Bewertung von *security* gibt es bewährte Kriterien, wie etwa die Common Criteria for Security Evaluation [8], die eine international anerkannte Zertifizierung der Sicherheitsleistung von KI-Technologie ermöglichen.

Die Kategorie *privacy* wird vereinzelt in technischen Standards adressiert, etwa im Zusammenhang mit der Fernauslesung von Stromzählern (Smart Metering). Die Abdeckung ist dabei regelmäßig lückenhaft, wie etwa die Beschränkung auf Pseudonymisierung und Vernachlässigung der Interventionsbefähigung. Für die Bewertung hat die Europäische Datenschutzgrundverordnung [9] einen Rahmen der Zertifizierung geschaffen. Die nötigen Kriterien und Infrastrukturen sind aber noch nicht vollständig verfügbar.

Für die Kategorie *autonomy* sind derzeit keine technischen Standards bekannt. Aktivitäten in diese Richtung werden aber insbesondere in der Normungsroadmap für KI [10] adressiert, deren Veröffentlichung für Ende 2020 geplant ist. In diesem Zusammenhang ist auch die Entwicklung von Bewertungskriterien anzustreben. Sie spielen eine zentrale Rolle für die Gestaltung von KI für den Menschen mit eingebauter digitaler Souveränität (Sovereignty by Design).

5 Reflexion

Wir haben den Zusammenhang von digitaler Souveränität und KI für den einzelnen Menschen untersucht. Dies betrifft alle Lebens- und Arbeitsbereiche aus der Perspektive der von den Auswirkungen von KI-Technologie betroffenen Personen. Unberücksichtigt bleibt die Perspektive von Unternehmen, Organisationen und staatlichen Organen. Dies bleibt einer weitergehenden Betrachtung vorbehalten.

Das wesentliche Ergebnis dieses Beitrags ist ein konkreter Vorschlag für Kategorisierung von digitaler Souveränität für den einzelnen Menschen mit Blick auf die Gestaltung von KI-Technologie. Eine solche Kategorisierung wird nach unserer Überzeugung dringend benötigt. Der Vorschlag sollte deshalb breit und gerne auch kontrovers diskutiert werden.

Literatur

1. Confederation of laboratories for artificial intelligence research in Europe (CLAIRE). https://claire-ai.org/. Zugegriffen: 15. Aug. 2020
2. AI ethics guidelines global inventory. https://inventory.algorithmwatch.org/. Zugegriffen: 15. Aug. 2020

3. European commission: ethics guidelines for trustworthy AI (2019). https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines. Zugegriffen: 15. Aug. 2020
4. Charta der Grundrechte der Europäischen Union. https://eur-lex.europa.eu/legal-content/DE/TXT/?uri=CELEX%3A12012P%2FTXT. Zugegriffen: 15. Aug. 2020
5. Das Standard-Datenschutzmodell, Version 2.0b (2020). https://www.bfdi.bund.de/DE/Datenschutz/Themen/Technische_Anwendungen/TechnischeAnwendungenArtikel/Standard-Datenschutzmodell.html. Zugegriffen: 15. Aug. 2020
6. Security by design principles according to OWASP. https://blog.threatpress.com/security-design-principles-owasp/. Zugegriffen: 15. Aug. 2020
7. European Union Agency for Cybersecurity (ENISA): Privacy and data protection by design (2015). https://www.enisa.europa.eu/publications/privacy-and-data-protection-by-design. Zugegriffen: 15. Aug. 2020
8. Common criteria for IT security evaluation, Version 3.1 Revision 5, April 2017. https://www.commoncriteriaportal.org/cc/. Zugegriffen: 15. Aug. 2020
9. Verordnung (EU) 2016/679 des Europäischen Parlaments und des Rates zum Schutz natürlicher Personen bei der Verarbeitung personenbezogener Daten, zum freien Datenverkehr und zur Aufhebung der Richtlinie 95/46/EG (Datenschutz-Grundverordnung) (2016). https://eur-lex.europa.eu/legal-content/DE/TXT/HTML/?uri=CELEX:32016R0679. Zugegriffen: 15. Aug. 2020
10. Normungsroadmap für KI. https://www.din.de/de/forschung-und-innovation/themen/kuenstliche-intelligenz/normungsroadmap-ki. Zugegriffen: 15. Aug. 2020