Kommentar

https://doi.org/10.5771/9783748913160-289 https://www.inlibra.com/de/page/agb - Open Access - (cc) BY

https://doi.org/10.5771/9783748913160-289 https://www.inlibra.com/de/page/agb - Open Access - (cc) BY

Sebastian Hallensleben, Andreas Hauschke

Zur Rolle von *Reliability* im Rahmen der Operationalisierung von KI-Ethik

Die Grundprinzipien für die ethischen Anforderungen an Künstliche Intelligenz sind erstaunlich unkontrovers. In einer Reihe vergleichender Studien[1] zeigt sich, dass Institutionen aus aller Welt[2] – darunter aus Europa, aus Amerika und auch aus Asien – ähnliche Prinzipien in den Vordergrund stellen. Dazu gehören beispielsweise *Transparency*, *Privacy*, *Sustainability*, *Accountability*, *Explainability*, *Fairness* und auch *Reliability*. (Wir verwenden hier bewusst die englischen Ausdrücke, um zusätzliche Unschärfen bei der Diskussion in mehreren Sprachen zu vermeiden.)

Während bei den Grundprinzipien oder ›Grundwerten‹[3] also eine gewisse Reife der globalen Diskussion konstatiert werden kann, steckt die Operationalisierung dieser Grundprinzipien noch in einem viel früheren Stadium. Aber ohne Operationalisierung bleiben die Prinzipien weitgehend wirkungslos: Softwareentwickler bleiben ratlos, wie sie den Code für ihre KI-Systeme konkret gestalten müssen, damit er etwa zu *Transparency* oder *Fairness* führt; Anbieter dieser Systeme kommen nicht über blumige Versprechungen im Stile überkommener Nachhaltigkeitsbehauptungen hinaus und für Kunden oder Anwender von KI fehlen handfeste und geprüfte oder zumindest prüfbare Systemeigenschaften, an denen sich die Verwirklichung ethischer Ansprüche festmachen ließe.

Kurz: Ethische Grundprinzipien müssen operationalisiert werden, beispielsweise durch eine methodisch bewährte Beschreibung mittels Kriterien, die ihrerseits in Form von Indikatoren und Observablen konkretisiert und messbar gemacht werden. Erst aus Messbarkeit und Prüfbarkeit erwächst eine Durchsetzbarkeit der Grundprinzipien – sei es in harter Form durch Regulierung für besonders risikobehaftete

1 Z.B. Thilo Hagendorff: »The Ethics of AI Ethics. An Evaluation of Guidelines«, in: *Minds and Machines* 30 (2020), S. 99–120.

2 Darunter viele Beratungsgremien, die einzelnen Regierungen oder Verbänden nahestehen, aber auch multinationale Organisationen (z.B. UNESCO COMIST), Standardisierungsorganisationen (z.B. IEEE), zivilgesellschaftliche Einrichtungen und auch einige große Firmen, insbesondere aus dem industriellen Bereich.

3 In der akademischen Diskussion um eine KI-Ethik wird oftmals die Fokussierung der Guidelines und der Standardisierungsbemühungen auf Werte als eine ›Verkürzung‹ oder ›Engführung‹ kritisiert. Das ist begreiflich angesichts der Kontroversen um den Status von Werten überhaupt. Folgt man hingegen dem pragmatischen Vorschlag von Victor Kraft, Werte als »Direktiven für das Verhalten«, als implizite Normen zu begreifen, bieten sie als (Grund-)Prinzipien die Basis für die Erarbeitung konkreterer Normierungen und Normordnungen (Victor Kraft: *Die Grundlagen einer wissenschaftlichen Wertlehre*, Wien 1951, S. 199).

https://doi.org/10.5771/9783748913160-289 https://www.inlibra.com/de/page/agb - Open Access - (cc) BY

Anwendungen oder in weicher Form durch den Wettbewerb im Markt. Erst aus Messbarkeit und Prüfbarkeit erwächst auch die Möglichkeit einer sichtbaren kontinuierlichen Verbesserung der Umsetzung ethischer Grundprinzipien.

Das Prinzip *Reliability* hat bei der Operationalisierung von KI-Ethik in zweierlei Hinsicht einen besonderen Status. Zum einen wird bisweilen kontrovers diskutiert, inwieweit *Reliability* überhaupt im Zusammenhang mit KI-Ethik diskutiert werden sollte, da *Reliability* doch auch als ein rein technischer Begriff verstanden werden könnte. Zum anderen spielt *Reliability* in der Gruppe der Grundprinzipien eine fundamentale Rolle und ermöglicht überhaupt erst die anderen Prinzipien. Damit soll allerdings keine hierarchische Wertung der Bedeutung angedeutet werden. Sind *Fairness* oder *Accountability* nicht mindestens genauso wichtig wie *Reliability*? Ist *Sustainability* angesichts der globalen Klimakrise nicht das wichtigste Prinzip überhaupt? Mag sein, und das soll hier weder infrage gestellt noch diskutiert werden. Aber *Reliability* ist die Voraussetzung dafür, dass ein KI-System in seiner technischen Realisierung die anderen Prinzipien überhaupt abbilden kann – und zwar zuverlässig, dauerhaft und nicht nur unter Schönwetterbedingungen.

Zunächst aber: Was ist im Zusammenhang mit KI-Ethik unter *Reliability* zu verstehen? Worin bestehen die entscheidenden Merkmale dieses abstrakten Prinzips und wie können diese eine Entscheidungshilfe bieten, wenn es um die Anwendung auf ein konkretes Gegenstandsfeld und die Beurteilung von Lösungen geht? Im Rahmen der Operationalisierung von KI-Ethik durch ein Werte-Kriterien-Indikatoren-Observablen-Modell (WKIO) und auf Grundlage aktueller Arbeiten der AI Ethics Impact Group[4] umfasst *Reliability* die folgenden vier Kriterien für ein KI-System:[5]

1. *Robustness*: Dieses Kriterium zielt auf das zuverlässige Funktionieren des Systems im gesamten Spektrum der im normalen Betrieb zu erwartenden Parameterkombinationen ab. Ist das System beispielsweise in der Lage, unterschiedliche Betriebsumgebungen zu erkennen? Ist das System rigoros getestet worden? Ist dabei der Grad der Korrektheit und Genauigkeit des Systemoutputs festgestellt worden und stehen die Metriken, die zur Feststellung der Korrektheit und Genauigkeit herangezogen werden, auf solidem Boden? Liefert das System bei vergleichbaren Eingaben auch immer vergleichbare Ausgaben? Und ist sichergestellt worden, dass möglichen systematischen Fehlern in den Eingaben

4 Vgl. AI Ethics Impact Group (Hg.): »Wie sich ethische Prinzipien für Künstliche Intelligenz in die Praxis übertragen lassen«, in: *AIEIG*, 11.06.2020, www.ai-ethics-impact.org (aufgerufen: 23.11.2021); das dort veröffentlichte Papier enthält auch eine Erläuterung des WKIO-Ansatzes, siehe Kapitel 2 in AI Ethics Impact Group (Hg.): »From Principles to Practice. An interdisciplinary framework to operationalise AI ethics«, in: *AIEIG*, 01.04.2020, https://www.ai-ethics-impact.org/resource/blob/1961130/c6db9894ee73aefa489d6249f5ee2b9f/aieig---report---download-hb-data.pdf (aufgerufen: 23.11.2021).

5 In Abwandlung des WKIO-Ansatzes wurde die Indikatorenebene hier in Form von Leitfragen formuliert. Dies hat den Vorteil, bereits einen ersten Schritt in Richtung der Erstellung von Prüfkatalogen zu gehen.

https://doi.org/10.5771/9783748913160-289 https://www.inlibra.com/de/page/agb - Open Access - (cc) BY

vorgebeugt wird? Wurde insgesamt verstanden und dokumentiert, in welchem Umfeld das System tatsächlich zuverlässig eingesetzt werden kann?

2. *Resilience*: Dieses Kriterium zielt auf die Zuverlässigkeit eines Systems in unerwarteten Situationen ab, die im normalen Betrieb eigentlich nicht auftreten sollten. Gibt es im System Fähigkeit und Kapazität zur Prädiktion und Vorbeugung des Auftretens solcher Situationen? Gibt es Reservesysteme? Gibt es eingebaute Redundanzen? Inwieweit kann das System mit fehlerhaften Daten umgehen, beziehungsweise erkennt es zumindest das Vorliegen fehlerhafter Eingaben? Wurde eine detaillierte Analyse möglicher Bedrohungen des Systems durchgeführt? Und wenn ja, ist hierfür eine ausgereifte und standardisierte Methodik verwendet worden?
3. *Cybersecurity*: Dieses Kriterium zielt speziell auf die IT-Sicherheit des Systems beziehungsweise die Reaktion des Systems auf Hackerangriffe ab. Dazu gehören nicht nur die Maßnahmen gegen unautorisierten Zugriff, sondern auch weitergehende Anforderungen, die sich auf Genese und Betriebsumfeld des Systems beziehen: Inwieweit ist garantiert, dass der genutzte Algorithmus nicht kompromittiert wurde? Inwieweit ist sichergestellt, dass die genutzten Daten nicht kompromittiert wurden, wobei die gesamte Lieferkette der Daten zu berücksichtigen ist? Gibt es ein etabliertes und möglichst standardisiertes Vorgehen zum Management von IT-Sicherheit für das System und für die Netzwerke, in die es eingebunden ist?
4. *Maintainability*: Dieses Kriterium zielt darauf ab, inwieweit ein System durch den Betreiber bzw. das Bedienpersonal gewartet und in Schuss gehalten werden kann. Sind für diese Zwecke ausreichend Informationen über das System verfügbar? Ist der Systemzustand jederzeit erkennbar und verständlich? Lässt sich das System aktualisieren, um beispielsweise Sicherheitslücken zu schließen oder auf veränderte regulatorische Rahmenbedingungen zu reagieren? Ist dabei sichergestellt, dass die Funktionen und andere wünschenswerte Eigenschaften des Systems nicht beeinträchtigt werden, und ist dies für den Betreiber auch ersichtlich? Werden Zwischenfälle aufgezeichnet und bei Bedarf beispielsweise an Aufsichtsbehörden gemeldet?

Die oben genannten Kriterien überlappen in der Welt der technischen Qualitätssicherung mit bestehenden Frameworks; sie sind also in der Technik nicht grundsätzlich neu. So wird beispielsweise eine zielführende und in der Praxis bewährte Handhabung von *Cybersecurity* je nach Anwendungsbereich in Standards wie IEC 62443 oder ISO 27000 beschrieben. Die Kriterien *Robustness* und *Resilience* sind in weiten Teilen durch Standards abgedeckt, die gemeinhin unter dem Begriff ›funktionale Sicherheit‹ zusammengefasst und diskutiert werden. Dazu gehört beispielsweise die Standardfamilie IEC 61508 mit ihren anwendungsspezifischen Ausdifferenzierun-

https://doi.org/10.5771/9783748913160-289 https://www.inlibra.com/de/page/agb - Open Access - (cc) BY

gen. Und das Kriterium *Maintainability* erinnert in vielerlei Hinsicht an EU-Regulierungen beispielsweise für Produktsicherheit.

Im Übrigen haben die Arbeiten der letzten beiden Jahre in Standardisierungsgremien auf europäischer Ebene (CEN-CENELEC AI Focus Group[6]) und internationaler Ebene (IEC SEG 10 »Ethics in Autonomous and Artificial Intelligence Applications«) gezeigt, dass sich viele Aspekte von KI-Ethik mit den etablierten Mitteln der technischen Standardisierung überraschend gut beschreiben und operationalisieren lassen. Dazu hat unter anderem eine klare Unterscheidung von Systembeschreibung und Werturteilen beigetragen, aber auch die Tatsache, dass Standardisierung von jeher als Konsensbildung unter allen betroffenen Stakeholdern (im Jargon: ›interessierten Kreisen‹) gelebt wird, mithin also in ihren Prozessen und Strukturen durchaus in der Lage ist, neben ingenieurtechnischen und naturwissenschaftlichen auch zivilgesellschaftliche und geisteswissenschaftliche Perspektiven substanziell einzubinden. Entgegen verbreiteter Klischees ist es nicht so, dass in der Standardisierung ausschließlich schmalspurige Techniker aktiv sind, denen ethische Fragestellungen weder bewusst noch wichtig sind. Vielmehr bilden Standardisierungsgremien ein breites Spektrum von Sichtweisen ab, und dies gilt insbesondere in der KI-Standardisierung. Wenn man eine zu füllende Lücke nennen möchte, dann wäre es eher die Expertise im sprachlich, inhaltlich und ideengeschichtlich präzisen Umgang mit Ethik, für die eine stärkere Mitwirkung beispielsweise von Vertretern der Technikphilosophie oder der Technikfolgenabschätzung wünschenswert wäre und die von den Normungsorganisationen bewusst gefördert werden sollte. Dies mag kurzfristig und vordergründig zu Verständigungsproblemen innerhalb von Standardisierungsgremien führen, da unterschiedliche Kulturen und Begriffswelten aufeinanderstoßen, aber allein schon die Einigung auf eine gemeinsame klar definierte Terminologie ist ein sowohl typisches als auch essenzielles Ergebnis von Standardisierungsarbeit und trägt nicht unerheblich zu ihrem Erfolg bei. Standardisierung ist nicht die willkürliche Festlegung von Regeln, sondern vielmehr das Ringen um einen breiten – im Fall von KI nicht nur technischen, sondern auch gesellschaftlichen – Konsens sowie dessen konkreter Formulierung in technisch nutzbarer Form.

Die oben am Beispiel *Reliability* erwähnte Möglichkeit des Verweises auf bestehende Standards und Regelwerke erleichtert es in der Praxis deutlich, KI-Ethik umzusetzen und zu prüfen: Die Einhaltung bestehender Standards liefert als Doppelfunktion einen Baustein für den Nachweis KI-ethischer Orientierung als Anerkennung notwendiger Bedingungen ihrer Umsetzung überhaupt sowie als konkrete Basis eines weiteren zu entwickelnden Umsetzungs- und Prüfgeschehens. Um es klar zu sagen: Die bestehenden Standards decken keineswegs sämtliche Aspekte

6 Vgl. Kapitel 5 in CEN-CENELEC Focus Group (Hg.): »Road Map on Artificial Intelligence (AI)«, in: *StandICT.eu*, September 2020, https://www.standict.eu/sites/default/files/2021-03/CEN-CLC_FGR_RoadMapAI.pdf (aufgerufen: 23.11.2021).

https://doi.org/10.5771/9783748913160-289 https://www.inlibra.com/de/page/agb - Open Access - (cc) BY

ab, die für KI-Ethik relevant sind. Sie bieten jedoch einen etablierten, in einem breiten Konsens entstandenen Sockel, auf den aufgebaut werden kann. Mit einer pragmatischen Nutzung des bereits über Jahrzehnte Entstandenen ist der Sache einer raschen und substanziellen Operationalisierung von KI-Ethik allemal mehr geholfen als mit einer fortdauernden Grundsatzdiskussion, die in Teilen noch nicht einmal bis zur Neuerfindung des Rads vorgedrungen ist.

Nicht zu unterschätzen ist auch, dass die oben implizierte Überlappung von Ethik und Technik für Ingenieure, die sich bisher mit KI-Ethik nicht auseinandergesetzt haben, einen Zugang eröffnet. Sie trägt dazu bei, dass KI-Ethik nicht als etwas grundsätzlich Neues, Ärgerliches oder gar Bedrohliches wahrgenommen wird, sondern als eine pragmatische und notwendige Erweiterung des etablierten und gewohnten Denkens über die Qualität von Produkten und die dazu notwendigen Prozesse und Strukturen.

Besteht dann aber nicht die Gefahr, dass KI-Ethik auf das technisch Prüfbare reduziert wird? Verschließen wir damit nicht die Augen vor den technisch nicht auflösbaren Dilemmata im Bereich der KI-Ethik (Stichwort ›Trolley-Problem‹) und der Diskussion des Umgangs mit ihnen? – Drehen wir diese Frage doch um: Wäre es angemessen, KI-Entwickler mit solchen Dilemmata allein zu lassen, ohne Standards und Regeln als Leitlinie? Soll der einzelne Ingenieur wirklich persönlich festlegen müssen, wie ein akzeptables Systemverhalten in ethisch heiklen Situationen aussieht? Sollen Aufsichtsbehörden die Verwendung von KI in einem spezifischen Anwendungsfall erlauben oder verbieten dürfen, ohne dafür konkrete, nachvollziehbare und prüfbare Systemeigenschaften zugrunde zu legen, z.B. Systemleistungen zur Prädiktion, Vorbeugung oder weitestmöglichen Verhinderung der Entstehung moralischer Dilemmata (die per definitionem ohnehin nicht auflösbar sind) – siehe oben das Kriterium *Resilience*?

Grundsätzlicher formuliert: Ist es wirklich richtig, wenn eine Gesellschaft zwar lautstark die Erfüllung gewisser ethischer Grundprinzipien fordert, sich aber davor drückt, die Verantwortung für deren schwierige Operationalisierung und die unvermeidlichen Herausforderungen in der Prüfung und Durchsetzung zu übernehmen? – Es reicht nicht aus, einerseits einen Dialog zwischen Wissenschaft, Technik und Gesellschaft dazu zu fordern, andererseits aber zentrale bereits bestehende Foren und Ergebnisse solcher Dialoge (siehe oben: etablierte Standards aus der Technik, z.B. für funktionale Sicherheit) nicht zu nutzen. Es reicht auch nicht aus, den Schwarzen Peter entweder dem Betreiber oder Bediener des KI-Systems oder aber den Gerichten zuzuschieben, die unscharf (in ›Generalklauseln‹) formulierte Zielvorgaben, Wertungen und Wunschvorstellungen auf den Einzelfall anwenden müssten. Nötig ist vielmehr eine Weiterführung der Arbeiten in der Gesetzgebung (vgl. den AI Act, der in der Europäischen Union entwickelt wird), in der Standardisierung und natürlich auch im gesellschaftlichen Diskurs insgesamt, um nicht nur

https://doi.org/10.5771/9783748913160-289 https://www.inlibra.com/de/page/agb - Open Access - (cc) BY

Grundprinzipien, sondern auch Wege der Operationalisierung von KI-Ethik konkret und in einem möglichst breiten Konsens zu formulieren.

Hier schließt sich der Kreis zur Betonung von *Reliability*: *Reliability* schlägt die Brücke von der klassischen Ingenieursverantwortung für zuverlässige, sichere und damit auch gesellschaftlich akzeptable Technik zu den zusätzlichen, weitergehenden Anforderungen an Systeme Künstlicher Intelligenz, die unter der Überschrift ›KI-Ethik‹ gestellt werden. Hier ist es am einfachsten, die Erfahrungen früherer Diskussionen zum gesellschaftlichen Umgang mit Technik in den aktuellen KI-Diskurs einzubringen und dort geltend zu machen. Diese Brücke sollten wir stärker nutzen.

Reliability ist gleichzeitig die Basis für die weitergehenden Anforderungen und Prinzipien von KI-Ethik. Diese sind nur dann erfüllbar, wenn ein KI-System die Eigenschaften, die dadurch jeweils benannt werden, auch über einen längeren Zeitraum und in unterschiedlichen Betriebszuständen und unter unterschiedlichen Umständen beibehält. Ohne eine solche Stabilität und Zuverlässigkeit sind Behauptungen zu ethischen Eigenschaften beinahe wertlos, da sie allenfalls einem Blitzlicht in einer ganz bestimmten einmaligen Situation entsprechen. Damit wird *Reliability* zu einem (aber nicht dem alleinigen) Fundament der anderen Prinzipien. Ein solides Funktionieren von KI aus ingenieurtechnischer Sicht ist eine essenzielle Voraussetzung, ohne die die Diskussion über KI-Ethik und das Ringen um eine geeignete Operationalisierung im luftleeren Raum hängen bleibt.

Mit anderen Worten: *Reliability* ist sowohl Brücke als auch Fundament, damit die Operationalisierung von KI-Ethik rasch und im breiten Konsens gelingt.

https://doi.org/10.5771/9783748913160-289 https://www.inlibra.com/de/page/agb - Open Access - (cc) BY