INPUT-PAPIER

KI-Ethik: Vom Prinzip zur individuellen Verantwortung

Ein Input für den 2. Workshop der Roundtable-Reihe ethische KI-Entwicklung (RTeKI)
13. September 2022, Julia Meisner, Linda Schwarz (Gesellschaft für Informatik e.V.)

1 ORIENTIERUNG AN ETHISCHEN WERTEN STÄRKT INNOVATION

Bleiben ethische Prinzipien in der KI-Entwicklung unberücksichtigt, besteht das Risiko, dass sich die so entwickelten Anwendungen fehlerhaft verhalten. Werden entsprechend schlechte KI-Systeme angewendet, drohen nicht nur negative Konsequenzen für die Zivilgesellschaft und die Umwelt. Es sinkt auch die öffentliche Akzeptanz algorithmischer Systeme, wodurch sich diese schwieriger etablieren lassen und Innovationen blockiert werden. Um die Entwicklung innovativer, sicherer und anerkannter KI-Systeme voranzutreiben, sollten ethische Werte folglich als wichtige Faktoren im KI-Entwicklungsprozess integriert werden – eine Anforderung, die zunehmend unumstritten ist.

Nachdem sowohl auf Regierungsebene (z. B. Europäische Kommission [1]), von (Fach-)Verbänden (z. B. IEEE [2]), Forschung und Zivilgesellschaft (z. B. Women Leading in AI [3], Fraunhofer IAIS [4]), privatwirtschaftlichen Unternehmen (**Anlage**) und weiteren Akteuren zahlreiche Prinzipien für die Entwicklung und den Umgang mit KI-Systemen nach ethischen Werten zusammengestellt wurden, verdeutlicht sich jedoch, dass Prinzipien allein nicht ausreichen [5].

2 PRINZIPIEN MÜSSEN IN PRAKTISCHE PROZESSE ÜBERSETZT WERDEN

Um wirksam zu werden, müssen Prinzipien in praktische Prozesse übersetzt und in ihrer Umsetzung durch diese geleitet werden. Das umfasst zunächst die **Ableitung** konkreter **Systemanforderungen** aus abstrakten Zielen, wie beispielsweise:

- **„Gerechtigkeit“** (*Justice*) kann nur herrschen, wenn KI-Systeme Individuen nicht unzulässig diskriminieren und somit Fairness sichergestellt wird [6]. Dafür müssen Verzerrungen (Bias) in allen Entwicklungsstufen entdeckt und vermieden werden [7].
- **„Unschädlichkeit“** (*Non-Maleficence*) kann nur garantiert werden werden, wenn KI-Systeme den Schutz von Daten und der Privatsphäre über den gesamten Systemlebenszyklus hinweg gewährleisten, sodass keine schädlichen Konsequenzen durch missbräuchliche Datennutzung entstehen [6].

Bei der Ableitung von Systemanforderungen können **Toolkits** für ethisches und verantwortungsbewusstes Design herangezogen werden, die (Entwicklungs-)Teams mit partizipativen Reflexions- und Gestaltungsübungen dabei helfen, ethische Werte in den Entwicklungsprozess einfließen zu lassen [8, 9].

Anschließend sollten softwareentwickelnde Teams an allen Entwicklungsstufen Begutachtungs- bzw. **Assessmentverfahren** durchführen, mit denen die Einhaltung ethischer Kriterien überprüft wird. Ansätze hierfür liegen bereits in verschiedenen Formen vor:

— Gemäß der **Assessment List for Trustworthy AI** der Europäischen Kommission sollen verantwortliche Personen etwa überprüfen, ob das KI-Modell auf Basis personenbezogener Daten trainiert wurde und ob Mechanismen eingerichtet wurden, durch die datenschutzrechtliche Probleme gemeldet werden können [10].
— Im Rahmen des IEEE **CertifAIEd Programms** helfen externe Expert*innen softwareentwickelnden Unternehmen dabei, ethische Risiken durch die Überprüfung der Kriterien Transparenz, Rechenschaft, algorithmischem Bias und Datenschutz auszuschließen [11].
— Durch Initiativen wie dem **AI Blindspots Discovery Process** des Berkman Klein Centers können softwareentwickelnde Teams unbeabsichtigte negative Konsequenzen (Blindspots) eines KI-Systems über den gesamten Entwicklungsprozess selbst aufdecken und adressieren [12].
— Mittels dem im Projekt ExamAI – KI Testing und Auditing [13] entwickelten Ansatz, **Acceptance Test Driven Development** mit den aus dem Safety Engineering bekannten **Assurance Cases** zu kombinieren, können sich Unternehmen erst auf ethische Prüfkriterien verständigen, das KI-System dann auf diese prüfen und die Ergebnisse abschließend dokumentieren [14].

❸ ETHIK BRAUCHT SOZIALE AUSHANDLUNG

Selbst praktische Ansätze wie die Formulierung konkreter Systemanforderungen oder die Durchführung von Assessmentverfahren können eine erfolgreiche KI-Entwicklung nach ethischen Maßstäben allerdings nur dann ermöglichen, wenn eine Reihe zusätzlicher Bedingungen erfüllt ist: *Erstens* müssen beteiligte Entwickler*innen wissen, wie die Anforderungen technisch umzusetzen sind. *Zweitens* muss klar sein, wer an welchen Stellen des Entwicklungsprozesses verantwortlich für die Begutachtung ist - und anhand welcher Kriterien diese durchzuführen ist [5]. Dafür muss *drittens* auch ein **gemeinsames Verständnis** der angestrebten Werte und Ziele vorliegen. Ethik stellt hier besondere Herausforderungen. Im Unterschied zu Maßnahmen wie technischen Standards oder Erklärbarkeit bzw. Nachvollziehbarkeit [15] von algorithmischen Entscheidungen drückt sich Ethik vor allem im Handeln aus und kann nur schwer explizit artikuliert werden [16]. Deshalb ist erfolgreiche KI-Entwicklung nach ethischen Maßstäben auf **soziale (Aushandlungs-)Prozesse** angewiesen und kann nicht allein durch die Anwendung ausgewählter praktischer Methoden vollzogen werden [17].

4 PERSÖNLICHE INVOLVIERUNG FÖRDERT ETHISCHES VERANTWORTUNGSBEWUSSTSEIN

Statt zu versuchen, ethische Prinzipien **„top down“** seitens der Managementebene in Unternehmen zu etablieren, empfiehlt der Technologiephilosoph Mark Coeckelbergh daher einen **„bottom up“ Ansatz**, bei dem Repräsentant*innen möglichst aller Geschäftsbereiche softwareentwickelnder Unternehmen sowie aus Zielgruppen der KI-Anwendungen zu einem gemeinsamen Austauschprozess über KI-Ethik eingeladen werden [16]. Dabei sollten die unterschiedlichen Perspektiven auf und Bedarfe in der Umsetzung von KI-Ethik als fester Bestandteil der Unternehmenskultur etabliert, Umsetzungsprozesse festgelegt und schließlich alle Mitarbeitende für den Umgang mit Ethik in der KI-Entwicklung sensibilisiert werden. Gemäß Thilo Hagendorff, Experte für Technik- und KI-Ethik, verspricht dieses Vorgehen, das Gefühl persönlicher Involvierung und Verbundenheit mit dem Thema KI-Ethik zu stärken. Mitarbeitende würden dadurch nicht durch normativen Zwang, sondern **individuelles moralisches Verantwortungsbewusstsein** zu ethischen Überlegungen in der KI-Entwicklung motiviert – auch dann, wenn sie gar nicht unmittelbar für ein bestimmtes KI-System verantwortlich sind [5]. Ethische Prinzipien würden dadurch tendenziell nicht mehr als hinderlich und irrelevant begriffen, sondern als wertvolle Faktoren in der KI-Entwicklung akzeptiert. So stiege die Wahrscheinlichkeit, sie konsequent ab den ersten Schritten des Entwicklungsprozesses zu berücksichtigen, statt sie allenfalls nachträglich als kosmetisches Marketinginstrument auf bestehende Praktiken und Ergebnisse anzuwenden [18].

5 DENKANSTÖSSE

a) Welche Austauschprozesse/-formate zu KI-Ethik gibt es in meinem Unternehmen?
b) Wer ist daran beteiligt?
c) Wie sind Verantwortlichkeiten zur Berücksichtigung von Ethik in der KI-Entwicklung in meinem Unternehmen geregelt?

LITERATUR

[1] Europäische Kommission, Generaldirektion Kommunikationsnetze, Inhalte und Technologien, „Ethik-leitlinien für eine vertrauenswürdige KI“, 2019. *https://data.europa.eu/doi/10.2759/22710*.

[2] The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, „General Principles“, 2019. *https://ethicsinaction.ieee.org/wp-content/uploads/ead1e_general_principles.pdf* (Zugriff am 30.8.2022).

[3] Women Leading in AI, „10 Principles of Resposible AI“, 2019. *https://womenleadinginai.org/wp-content/uploads/2019/02/WLiAI-Report-2019.pdf* (Zugriff am 30.8.2022).

[4] Fraunhofer-Institut für Intelligente Analyse und Informationssysteme IAIS, „Vertrauenswürdiger Einsatz von Künstlicher Intelligenz“, 2019. *https://www.iais.fraunhofer.de/content/dam/iais/KINRW/Whitepaper_KI-Zertifizierung.pdf* (Zugriff am 30.8.2022).

[5] Hagendorff, T., „A Virtue-Based Framework to Support Putting AI Ethics into Practice“. Philosophy & Technology, Bd. 35, 2022. *https://doi.org/10.1007/s13347-022-00553-z.*

[6] Morley J., Floridi L., Kinsey L., Elhalal A., „From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices“, Science and Engineering Ethics, Bd. 26, Nr. 4, S. 2141-2168, 2020. *https://doi.org/10.1007/s11948-019-00165-5.*

[7] Hallensleben, S., Hustedt, C., Fetic, L., Fleischer, T., Grünke, P., Hagendorff, T., et al., „ From principles to practice: An interdisciplinary framework to operationalise AI ethics“. Gütersloh: Bertelsmann Stiftung, S. 1–56, 2019.

[8] Ethics Toolkit: *http://ethicskit.org/tools.html* (Zugriff am 30.8.2022).

[9] Responsible Tech Design Library: *https://www.responsibletechdesign.com* (Zugriff am 30.8.2022).

[10] European Commission, „Assessment List for Trustworthy AI (ALTAI)“, 2020. *https://futurium.ec.europa.eu/en/european-ai-alliance/pages/welcome-altai-portal* (Zugriff am 30.8.2022).

[11] IEEE CertifAIEd: *https://engagestandards.ieee.org/ieeecertifaied.html* (Zugriff am 30.8.2022).

[12] AI Blindspot: *https://aiblindspot.media.mit.edu* (Zugriff am 30.8.2022).

[13] Gesellschaft für Informatik e.V., „Abschlussbericht ExamAI – KI Testing und Auditing“, 2021. *https://testing-ai.gi.de/fileadmin/PR/Testing-AI/Abschlussbericht_ExamAI_-_KI_Testing_und_Auditing.pdf.*

[14] Hauer, M. P., Adler, R., & Zweig, K., „Assuring Fairness of Algorithmic Decision Making“, 2021 IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW), S. 110-113, 2021. *https://ieeexplore.ieee.org/document/9440188*.

[15] Henriksen, A., Enni, S., Bechmann, A., „Situated Accountability: Ethical Principles, Certification Standards, and Explanation Methods in Applied AI“. Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES ‘21). Association for Computing Machinery, New York, NY, S. 574–585, 2021. *https://doi.org/10.1145/3461702.3462564*.

[16] Coeckelbergh, M., „AI Ethics“, Cambridge, MA u. a.: The MIT Press, 2020.

[17] Theodoru, A., Dignum, V., „Towards ethical and socio-legal governance in AI“. Nature Machine Intelligence, Bd. 2, S. 10–12, 2020. *https://doi.org/10.1038/s42256-019-0136-y.*

[18] Floridi, L., „Translating Principles into Practices of Digital Ethics: Five Risks of Being Unethical“. Philosophy & Technology, Bd. 32, Nr. 2, 185–193, 2019. https://doi.org/10.1007/s13347-019-00354-x.

ANLAGE

Die teilnehmenden Unternehmen der Roundtable-Reihe haben eine große Breite an unternehmensinternen ethischen Prinzipien für die Entwicklung aufgestellt und diese zum Teil bereits veröffentlicht. Ordnet man diese technischen Systemanforderungen (technische Prinzipien) und sozialen Prozessen (soziale Prinzipien) zu, ergibt sich folgendes Bild. Auffällig ist, dass einige Prinzipien (z. B. Fairness) sowohl als Systemanforderung verstanden werden, als auch als etwas, das innerhalb einer Unternehmenskultur herzustellen und durch Mitarbeitende zu berücksichtigen ist.

Auch konkrete Prozesse und Verantwortlichkeiten zum Austausch über und zur Umsetzung von Ethikprinzipien finden sich bei einigen Unternehmen. Lücken bedeuten nicht, dass Prinzipien, Prozesse und Verantwortlichkeiten nicht bestehen, sondern dass dazu aktuell keine Informationen vorliegen.

UNTERNEHMEN	TECHNISCHE PRINZIPIEN	SOZIALE PRINZIPIEN	PROZESSE & VERANTWORTLICHKEITEN
ai-omatic	— Robustheit — Nachvollziehbarkeit — Vertrauenswürdigkeit — Berücksichtigung von Datenschutz und Privatsphäre	— Vorrang des menschlichen Handelns — Berücksichtigung rechtlicher und ethischer Grundsätze	
Aleph Alpha		— Neugierde, Wachstum und technologisches Verständnis — Fairness und Ausgewogenheit — Kollaboration und Gemeinschaft — Positive Auswirkungen	
BMW	— Robustheit und Sicherheit — Daten und Privatsphärenschutz — Transparenz und Erklärbarkeit	— Menschliche Aufsicht und Handlungsmacht — Diversität, Nichtdiskriminierung und Fairness — Wohlergehen von Umwelt und Gesellschaft	
Bundesagentur für Arbeit	— Robustheit, Sicherheit und Verlässlichkeit — Schutz der Privatsphäre und Datenqualität — Transparenz und Erklärbarkeit — Gerechtigkeit, Inklusion und Vielfalt	— Der Mensch im Mittelpunkt — Gesellschaftlicher Nutzen und Gemeinwohlorientierung — Schutz der Privatsphäre und Datenqualität — Transparenz und Erklärbarkeit — Gerechtigkeit, Inklusion und Vielfalt — Rechenschaftspflicht und Rechtmäßigkeit	

UNTERNEHMEN	TECHNISCHE PRINZIPIEN	SOZIALE PRINZIPIEN	PROZESSE & VERANTWORTLICHKEITEN
Continental	— Nachvollziehbarkeit — Datensicherheit	— Verantwortungsbewusstsein — Rechtmäßigkeit	
Deutsche Telekom	— Sicherheit — Verlässlichkeit — Vertrauen (Menschlicher Eingriff möglich)	— Verantwortung — Sorgsamkeit, Recht- und Gesetzmäßigkeit — Unterstützung (Kund*innen im Mittelpunkt) — Transparenz — Kooperation (Mensch – Maschine) — Teilen und Erklären	— Selbstverpflichtende Professionsethik legt konkrete To-dos und Verantwortlichkeiten für einzelne Prinzipien und Entwicklungsstufen fest
Lufthansa Industry Solutions	— Beachtung allgemeingültiger Vorschriften — Orientierung an menschlichen Rechten, Werten und Solidarität — Unterstützung und Erleichterung menschlichen Lebens, ohne menschliche Kontrolle auszuschließen — Verständlichkeit und Nachvollziehbarkeit — Algorithmische Rechenschaftspflicht — Menschen fair und diskriminierungsfrei behandeln — Menschen unterstützen und beteiligen — Sicherheit — Privatsphärenschutz — Zuverlässigkeit — Sicherheit		

UNTERNEHMEN	TECHNISCHE PRINZIPIEN	SOZIALE PRINZIPIEN	PROZESSE & VERANTWORTLICHKEITEN
Microsoft	— Fairness — Zuverlässigkeit und Sicherheit — Sicherheit und Datenschutz — Fairness — Inklusion — Transparenz	— Rechenschaftspflicht	— Responsible AI Standard definiert Anforderungen für verantwortungsvolle Produktentwicklung und leitet Teams bei ihrer Umsetzung an — Aether-Ausschuss berät Geschäftsführung zu Herausforderungen und Chancen von KI — Initiative und Entwicklungsteam RAISE (Responsible AI Strategy in Engineering)
ML6	— Robustheit — Sicherheit — Datenschutz — Förderung von Vielfalt, Fairness, Erklärbarkeit und Transparenz	— Gesellschaftlicher Nutzen — Verantwortlichkeit und Rechenschaftspflicht — Respekt vor menschlicher Autonomie	— Ethics Unit beschäftigt sich mit ethischen Risiken, KI-Prinzipien und regulatorischen Anforderungen und gibt Expertise intern und extern an Kund*innen weiter — Austausch mit privaten & öffentlichen Organisationen — Mitarbeitendentraining — Internes AI Ethical Advisory Board prüft sensible Projekte
SAP	— Transparenz — Integrität (zur vorgesehenen Nutzung) — Qualität und Sicherheit — Daten- und Privatsphärenschutz	— Werteorientierung an (Menschen-)Rechten und internationalen Normen — Inklusives Design — Bewusstsein und Vermeidung von Bias — Beschäftigung mit gesellschaftlichen Herausforderungen von KI	— Internes Ethics Steering Comittee aus Jurist*innen, Entwickler*innen, Personaler*innen behält Überblick über Aktivitäten und leitet Mitarbeitende an — AI Ethics Advisory Panel aus externen Expert*innen berät Committee fachlich — Entwicklung in Zusammenarbeit mit Nutzenden