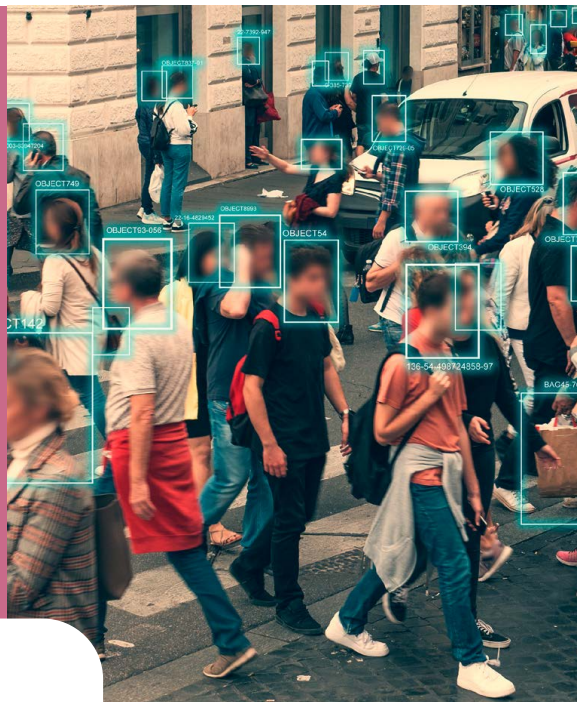
Lernende Systeme

DIE PLATTFORM FÜR KÜNSTLICHE INTELLIGENZ

Kritikalität von KI-Systemen in ihren jeweiligen Anwendungskontexten

Whitepaper von Jessica Heesen,
Jörn Müller-Quade, Stefan Wrobel et al.
AG IT-Sicherheit, Privacy, Recht und Ethik
AG Technologische Wegbereiter und
Data Science



Kurzfassung

Künstliche Intelligenz verbessert Prozesse und Geschäftsmodelle und hilft dabei, die Zukunftsfähigkeit von Wirtschaft und Gesellschaft zu sichern. So können KI-Systeme in der Medizin Diagnosen und Behandlungen verbessern, in der Mobilität die Routenplanung optimieren oder eine passgenauere Abstimmung zwischen Bedarfen und Angeboten ermöglichen. Gleichzeitig bergen sie aber auch Risiken und bringen damit das Thema Künstliche Intelligenz (KI) und Vertrauenswürdigkeit ins Spiel. Denn die Risiken, die der Einsatz von KI-Systemen mit sich bringen kann, und die Schäden, die bei diesen Einsätzen entstehen können, sind vielfältig und oft nur schwer abschätzbar.

Um sichere und zuverlässige Anwendungen in den Einsatz zu bringen, hat sich die Europäische Kommission in ihrem Vorschlag zur Regulierung von KI-Systemen im April 2021 dafür ausgesprochen, KI-Systeme entsprechend ihrem Gefahrenpotenzial zu regulieren und sie in vier Risikostufen zu klassifizieren, von minimalem Risiko (kein Regulierungsbedarf) bis hin zu inakzeptablem Risiko (Verbot der Anwendung). Mit ausgewählten Inhalten dieses EU-Vorschlags setzen sich Expertinnen und Experten der Arbeitsgruppen IT-Sicherheit, Privacy, Recht und Ethik sowie Technologische Wegbereiter und Data Science im Whitepaper „Kritikalität von KI-Systemen in ihren jeweiligen Anwendungskontexten – Ein notwendiger, aber nicht hinreichender Baustein für Vertrauenswürdigkeit“ auseinander: nämlich anhand welcher Kriterien festgelegt werden kann, in welchen Fällen der Einsatz von KI-Systemen von vornherein zu regulieren ist und wann dies nicht notwendig ist. Mit dieser zentralen Fragestellung als Leitgedanken möchten sie eine „gute“ Antwort geben, wie durch Regulierung KI-Qualität sichergestellt werden kann und wie Überregulierung vermieden und zugleich Innovationen gefördert werden können, um letztlich somit den Schutz des Subjekts zu gewährleisten. Damit bereichern sie die aktuelle politische Debatte um weitere Perspektiven, die das Thema Kritikalität von KI-Systemen hinsichtlich ihrer Vertrauenswürdigkeit aufnehmen.

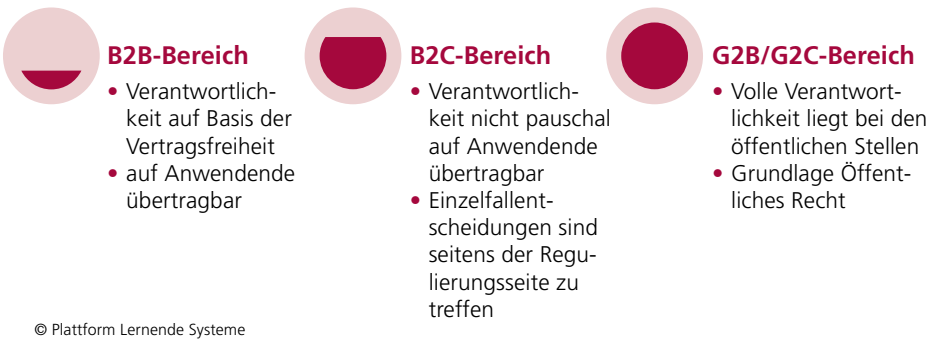
Ex-ante- und Ex-post-Maßnahmen

Grundsätzlich empfehlen die Autorinnen und Autoren die Kritikalitätsbetrachtung von KI-Systemen bereits im Vorhinein (ex ante) durch Maßnahmen zu ergänzen, die im Nachhinein greifen (Kapitel 1). Diese Maßnahmen im Nachhinein (ex post) können im Hinblick auf Transparenz und Nachvollziehbarkeit sowie Haftung und Schadensersatz nur dann erfolgversprechend sein, wenn der Grundstein hierfür im Vorhinein gelegt wurde. Als Beispiel werden effektive, niedrigschwellige und zeitnahe Beschwerde- und Verbraucherschutzregime aufgeführt. Denn diese stärken und sichern die Datensouveränität des Datensubjekts über die Einwilligung am Beginn des Datenverarbeitungsprozesses hinaus. Die Betrachtung der Kritikalität sollte daher im Sinne einer Gefahrenvorsorge statt einer Gefahrenabwehr erfolgen, da Risiken nicht nur rein technisch, sondern auch sozio-technisch zu bewerten sind.

Verantwortung, Verantwortungsketten und Haftung

Bei der Betrachtung der Kritikalität von KI-Systemen in bestimmten Anwendungskontexten sollten zum anderen auch unterschiedliche Verantwortungs- und Betroffenheitsdimensionen herangezogen werden (Kapitel 2.3). Diese Aufteilung der Verantwortung des Risikos über Haftungsregelungen bildet einen wichtigen Baustein hin zu vertrauenswürdigen KI-Systemen. Ziel der (Unternehmens-)Haftung ist es, das technische Risiko nach Gesichtspunkten der Fairness und Funktionalität unter den unterschiedlichen Akteuren bestmöglich zu teilen. Dabei geht es zentral um die Frage, wer ein Risiko besser einschätzen, besser tragen und gegebenenfalls beseitigen kann. Hierfür stellen die Autorinnen und Autoren ein Modell zur Verantwortungsaufteilung im Fall von Schäden durch KI-Systeme vor, das zum einen nach der Adressatin/dem Adressaten und zum anderen gegebenenfalls nach der Kritikalität des jeweiligen Anwendungsgebiets unterscheidet. So sollten bei Haftungsfragen im B2B-Bereich grundsätzlich die Bedingungen des Vertragsrechts gelten (außer in gesellschaftlich kritischen Bereichen), während im B2C-Bereich Entscheidungen in Abhängigkeit von der Kritikalität getroffen werden sollten. Im B2B-Bereich sollte grundsätzlich die Anwenderin oder der Anwender entsprechend dem Vertragsrecht verantwortlich für die von KI-Systemen gelieferten Ergebnisse sein. Für Anwendungen im öffentlichen Bereich sollte die öffentliche Hand die volle Verantwortung für diskriminierende oder schädliche Konsequenzen im Rahmen des öffentlichen Rechts tragen. Diese Überlegungen zur Verantwortung und auch Haftung sind nicht nur national, sondern auch europaweit in den Blick zu nehmen.

Abbildung 1: Verantwortungsketten nach Bereich



Dimensionen zur Bewertung: Kontroll- und Entscheidungsmöglichkeiten

Neben der Regulierung anhand der Kritikalität im Vorhinein, dem Schaffen von Verbraucherschutzregimen für mögliche Schadensfälle, die Aufteilung der Verantwortung des Risikos über Haftungsregelungen empfehlen die Autorinnen und Autoren auch stärker die Kontroll- und Entscheidungsmöglichkeiten der Nutzenden von KI-Systemen bei der Bewertung der Kritikalität in den Blick zu nehmen (Kapitel 3). Hierfür schlagen sie vor, die Kriterien zur Bewertung der Kritikalität eines KI-Systems in einem bestimmten Anwendungskontext in zwei Dimensionen einzuteilen. Nämlich, ob die Empfehlungen oder Entscheidungen eines KI-Systems Menschenleben oder Rechtsgüter wie die Umwelt gefährden und wieviel Handlungsspielraum dem Menschen bei der Auswahl und Nutzung der Anwendung bleibt, etwa um bestimmte Funktionen abzuschalten: Je höher das Ausmaß der möglichen Verletzung von Menschenleben und weiteren hohen Rechtsgütern und je geringer der Umfang der Handlungsmöglichkeiten des Individuums, desto höher ist die Kritikalität – und daraus abgeleitet der Regulierungsbedarf – und vice versa.

Abbildung 2: Kritikalität von KI-Systemen vor dem Hintergrund ihres jeweiligen Anwendungskontextes

Höhere Kritikalität

- Möglichkeit und Ausmaß der Verletzung von Menschenleben und weiteren hohen Rechtsgütern: **groß**
- Umfang der Handlungsmöglichkeiten der Individuen: **gering**

Niedrigere Kritikalität

- Möglichkeit und Ausmaß der Verletzung von Menschenleben und weiteren hohen Rechtsgütern: **gering**
- Umfang der Handlungsmöglichkeiten der Individuen: **groß**

Die vorangestellten Ausführungen und genannten Ansatzpunkten werden von Expertinnen und Experten der Plattform Lernende Systeme in kurzen Interviews um weitere Aspekte aus ihrer jeweiligen Fachexpertise heraus ergänzt und genauer erörtert (Kapitel 4). Dieser thematische Bogen bildet zum einen die Breite und zum anderen die Vielschichtigkeit von weiteren möglichen Antworten auf die Frage nach der Regulierung von KI-Systemen in Abhängigkeit von der Kritikalität ab: Wie hängen Kritikalität und Regulierung zusammen? Wie ist das Konzept der Kritikalität zu ver-

stehen? Oder: Welche ethischen Ansprüche ergeben sich hinsichtlich Verantwortung oder für das technischen Handeln? um nur einige der Fragestellungen zu nennen. In diesem Fragen-Antworten-Format spiegelt sich die Komplexität der Kritikalitätsbewertung von KI-Systemen deutlich wider und zeigt zugleich auf, dass das Thema noch nicht umfassend und abschließend zu betrachten ist.

Eine Zusammenfassung der Anpassungsvorschläge zur Präzisierung und Konkretisierung des EU-Regulierungsvorschlags seitens der Expertinnen und Experten gibt folgende Abbildung:

Abbildung 3: Anpassungsvorschläge zum Regulierungsvorschlag der Europäischen Kommission

EU-Regulierung	Anpassungsvorschläge
Möglichkeit und Ausmaß der Verletzung von Menschenleben und weiteren Rechtsgütern	
Schweregrad des Schadens • Ausmaß, in dem ein KI-System einen Schaden verursacht hat oder das Risiko eines Schadens besteht • Ausmaß der Auswirkungen des Schadens • Wahrscheinlichkeit, dass das KI-System eine hohe Anzahl von Personen schädigt • Wahrscheinlichkeit, dass ein KI-System mehr als einen der speziell definierten Schäden verursacht	**Problem:** schwierige Quantifizierbarkeit → Betrachtung gesamtgesellschaftlicher und individueller Schäden → Betrachtung materieller und immaterieller Schäden
Wahrscheinlichkeit eines Schadens • Anzahl der Nutzenden des KI-Systems +	• Anzahl der Nutzungen des KI-Systems durch die Nutzenden • Persistenz der Bedrohungssituation • Kontrollierbarkeit der Bedrohungssituation (Grad der Vernetztheit)
Umfang der Handlungsfreiheiten der Individuen	
• Ausmaß der Abhängigkeit der potenziell von einem Ergebnis betroffenen Personen • Ausmaß der Vulnerabilität der potenziell betroffenen Personen gegenüber dem/der Nutzenden eines KI-Systems • Ausmaß der Reversibilität des von einem KI-System erzeugten Ergebnisses • Verfügbarkeit und Wirksamkeit von Rechtsmitteln (im Unionsrecht und im Recht der Mitgliedstaaten) • Ausmaß, in dem die bestehenden Rechtsvorschriften der Union in der Lage sind, die vom KI-System ausgehenden Risiken zu verhindern/zu minimieren +	• Opt-out-/Konfigurationsmöglichkeiten der Nutzenden • Entscheidungsmöglichkeiten der Nutzenden • Marktstruktur/Pluralität des Dienstangebots

Die vorliegende Analyse zeigt, dass das von der Europäischen Kommission gewählte Konzept der Risiko- oder auch Kritikalitätsbewertung zur Sicherstellung der Qualität von KI-Systemen eine gute Orientierungsfunktion zur Bewertung und Regulierung hat – wenn das Konzept um einige weitere Einordnungskriterien ergänzt wird. Nach Ansicht der Expertinnen und Experten ist es unerlässlich, das Konzept zu erweitern und weitere Kriterien zu definieren sowie zu präzisieren, um das Gefahrenpotenzial eines KI-Systems bemessen und beurteilen zu können. Vor allem, so betonen sie, ist es entscheidend, dass die Regulierung von KI-Systemen immer vor dem Hintergrund des jeweiligen Anwendungskontextes zu sehen ist. Dies auch, um damit eine Balance zwischen Innovationsoffenheit auf der einen und Schutz des Subjekts auf der anderen Seite zu schaffen.

Neben der Regulierung anhand der Kritikalität im Vorhinein, dem Schaffen von Verbraucherschutzregimen für mögliche Schadensfälle, die Aufteilung der Verantwortung des Risikos über Haftungsregelungen bilden auch die Kontroll- und Entscheidungsmöglichkeiten der Nutzerinnen und Nutzer von KI-Systemen bei der Bewertung der Kritikalität einen wichtigen – wenn auch nicht hinreichenden – Baustein hin zu vertrauenswürdigen KI-Systemen. Denn gleichzeitig sehen die Autorinnen und Autoren noch weiteren Forschungsbedarf zur Überwindung von Schwachstellen (Komplexitätsreduktion, mangelnde Vorherseh- und Absehbarkeit von Schäden), die es gilt auch auf in Zukunft mit in den Blick zu nehmen.

Impressum
Herausgeber: Lernende Systeme – Die Plattform für Künstliche Intelligenz | Geschäftsstelle | c/o acatech | Karolinenplatz 4 | D-80333 München | kontakt@plattform-lernende-systeme.de | www.plattform-lernende-systeme.de |
Folgen Sie uns auf Twitter: @LernendeSysteme | Stand: November 2021 | Bildnachweis: DedMityay/shutterstock/Titel

Diese Kurzfassung entstand auf Grundlage des Whitepapers *Kritikalität von KI-Systemen in ihren jeweiligen Anwendungskontexten – Ein notwendiger, aber nicht hinreichender Baustein für Vertrauenswürdigkeit.* München, 2021. Es wurde erstellt von Mitgliedern der Arbeitsgruppen IT-Sicherheit, Privacy, Recht und Ethik sowie Technologische Wegbereiter und Data Science. Die Originalfassung der Publikation ist online verfügbar unter: https://www.plattform-lernende-systeme.de/publikationen.html

GEFÖRDERT VOM

Bundesministerium für Bildung und Forschung

acatech
DEUTSCHE AKADEMIE DER TECHNIKWISSENSCHAFTEN