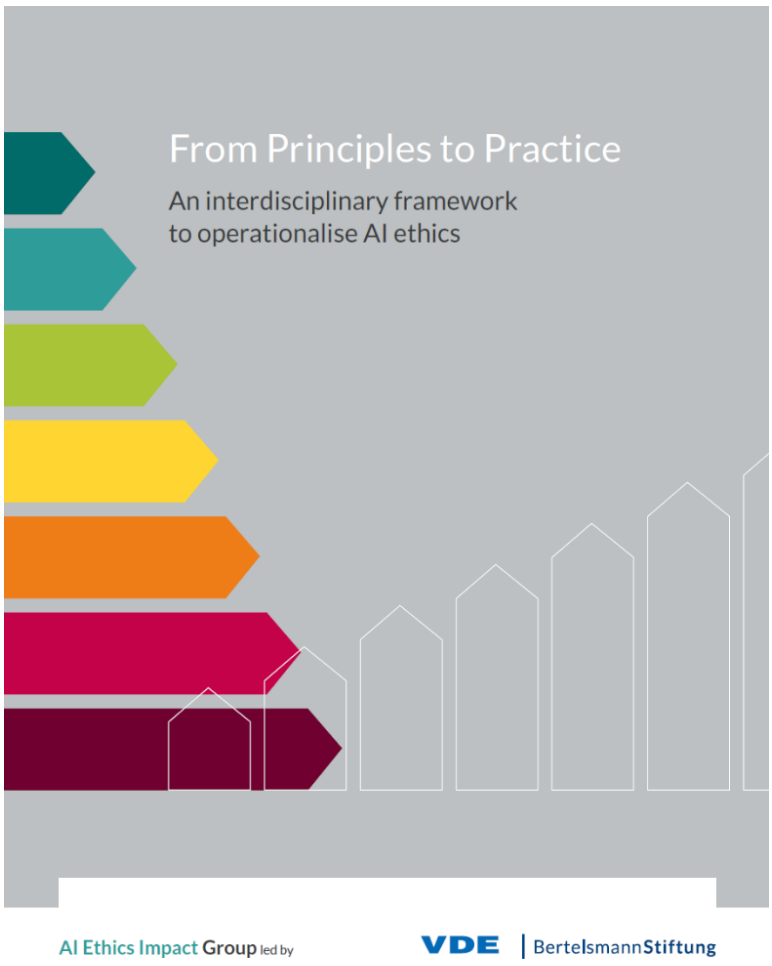
Wie sich ethische Prinzipien für Künstliche Intelligenz in die Praxis übertragen lassen

Deutsche Kurzfassung des Reports

„From Principles to Practice – An interdisciplinary framework to operationalise AI ethics"

Juni 2020



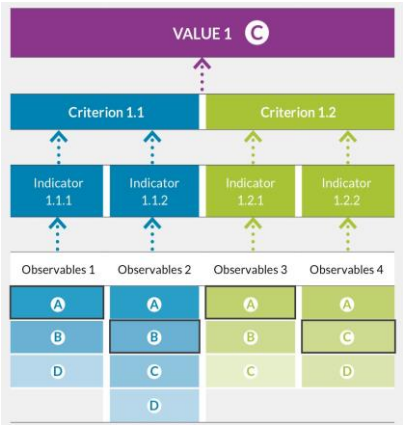
Zeitlich abgestimmt auf das Erscheinen des White Papers der EU-Kommission hat die AI Ethics Impact Group ein Konzept ausgearbeitet, wie sich KI-Ethik in die Praxis bringen lässt. Einige der Grundgedanken wurden vom VDE e.V. und der Bertelsmann Stiftung bereits Anfang Mai 2019 bei einer Experten-anhörung der Enquetekom-mission Künstliche Intelligenz vorgestellt. Der folgende Text ist eine deutsche Kurzfassung der im April 2020 vorgestellten englischen Vollversion. (ver-fügbar unter www.ai-ethics-impact.org).

Kontakt: sebastian.hallensleben@vde.com und carla.hustedt@bertelsmann-stiftung.de

Überblick: Die drei Elemente des KI-Ethik-Frameworks

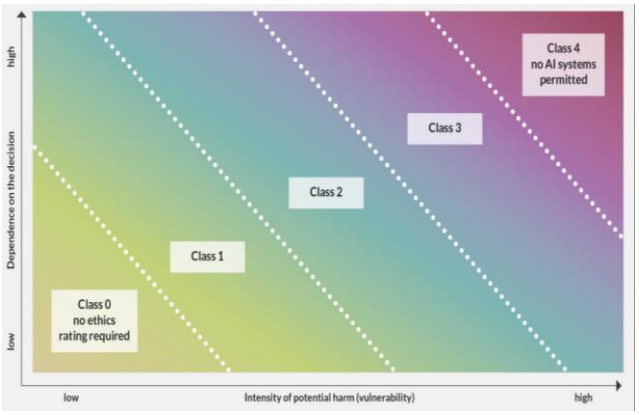
Verschiedene Interessengruppen haben Richtlinien für KI-Ethik vorgelegt, die als Grundlage für politische Bewertungen herangezogen werden. Fast alle erkennen ähnliche Werte als entscheidend und als Mindestanforderung für ethische KI-Anwendungen an - einschließlich Datenschutz, Fairness oder Nichtdiskriminierung sowie Transparenz und Sicherheit. Es bleibt jedoch unklar, wie die praktische Umsetzung dieser Grundsätze gewährleistet werden und mit der technischen und gesellschaftlichen Entwicklung Schritt halten kann. Dieses Fehlen spezifischer und überprüfbarer Prinzipien gefährdet die Wirksamkeit und Durchsetzbarkeit ethischer Richtlinien für automatisierte Entscheidungssysteme („algorithmic decision-making systems, ADM"). Wir schlagen vor, diese Lücke durch die Kombination dreier Elemente zu schließen, die nachstehend im Überblick skizziert und in den folgenden Kapiteln hergeleitet und näher erläutert werden:

1) Ein detailliertes KI-Ethik-Label (Abb. rechts), das dem Energieeffizienzlabel nachempfunden ist und so auf einen Blick Orientierung und kompakte Informationen zu den ethisch relevanten Eigenschaften eines KI-Systems bietet. Erreicht werden kann damit eine bessere Kontrolle durch politische Entscheidungsträger*innen, Regulierungsbehörden, Aufsichtsgremien, Überwachungs- und normgebende Organisationen. Gleichzeitig bietet eine solche Kennzeichnung auch Entwicklern Klarheit, die versuchen, ethisch einwandfreie KI-Systeme zu schaffen, und erhöht zugleich die Transparenz und Vergleichbarkeit von Produkten für die Nutzer. Darüber hinaus schafft solch ein detailliertes und aussagekräftiges Label Markttransparenz sowohl in B2B- als auch B2B-Kontexten.



2) Ein WKIO-Modell (engl.: VCIO für Values, Criteria, Indicators, Observables) als Ansatz für die Konkretisierung und Messbarmachung der im KI-Ethik-Label aufgeführten Werte. Die Methode hilft dabei, ethische Prinzipien umsetzbar, vergleichbar und messbar zu machen. Dabei werden den Werten zu bestimmende (variable) Kriterien zugeordnet, deren Bewertung anhand von Indikatoren und Observablen (messbar) erfolgt und so den Grad der jeweiligen Werterfüllung anzeigt. Werte werden so differenziert betrachtet und bewertet und müssen nicht absolut bestimmt werden. Das Modell lässt Raum für eine Differenzierung der KI-Systeme in ihren verschiedenen Anwendungskontexten.

3) Eine Risikomatrix, ein zweidimensionales Modell für die Klassifizierung der verschiedenen Anwendungskontexte von KI-Systemen. Diese abgestufte Bewertung der Kritikalität automatisierter Entscheidungssysteme in Abhängigkeit des Einsatzbereiches der KI, beugt zudem der Möglichkeit der Über- oder Unterregulierung vor. Der risikobasierte Ansatz, der ursprünglich an der Universität Kaiserslautern um KI-Forscherin Katharina Zweig entwickelt wurde, wurde auch von der Datenethikkommission der Bundesregierung aufgegriffen.

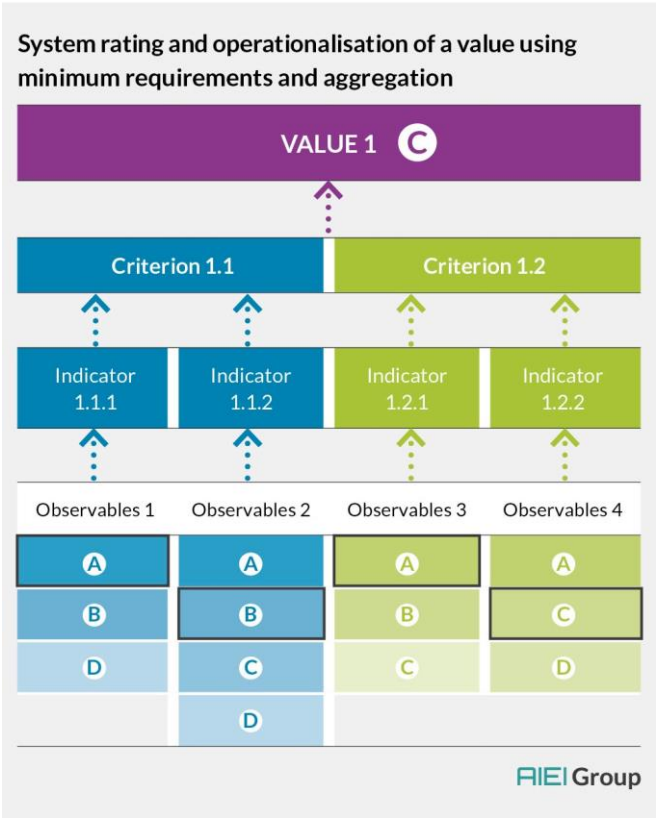


Politische Entscheidungsträger*innen und Regulierungsbehörden können die Kombination aus Risikomatrix, WKIO-Modell und Ethik-Label nutzen, um Anforderungen für verschiedene Anwendungskontexte zu spezifizieren und eine Überregulierung von Anwendungsbereichen zu vermeiden, die keine wesentlichen ethischen Herausforderungen darstellen. Für Anwendungsfelder, die in eine der höheren Risikostufen eingestuft werden, können sie verlangen, dass hier ein algorithmisches Entscheidungssystem (1) ein Ethik-Label tragen muss, welches den Grad der Werterfüllung für Werte wie bspw. Transparenz, Robustheit oder Gerechtigkeit angibt, und (2) gewisse Mindestniveaus innerhalb dieses Labels gewährleistet sein müssen.

Der WKIO-Ansatz und das KI-Ethik Label im Detail

Um zu einem Label zu kommen, bedarf es der methodischen und nachvollziehbaren Bewertung ethischer Prinzipien. Dies leistet der WKIO-Ansatz, der (1) klärt, was mit einem bestimmten Wert gemeint ist (Wertdefinition), dann (2) erklärt, wie zu prüfen oder zu beobachten ist, ob oder inwieweit ein technisches System einen Wert erfüllt oder verletzt (Messung) und (3) die Existenz von Wertkonflikten anerkennt und angibt, wie mit diesen Konflikten im jeweiligen Anwendungskontext der KI umzugehen ist.

>> Zur praktischen Umsetzung der KI-Ethik operiert der WKIO-Ansatz also auf vier Ebenen: Werte, Kriterien, Indikatoren & Observablen.

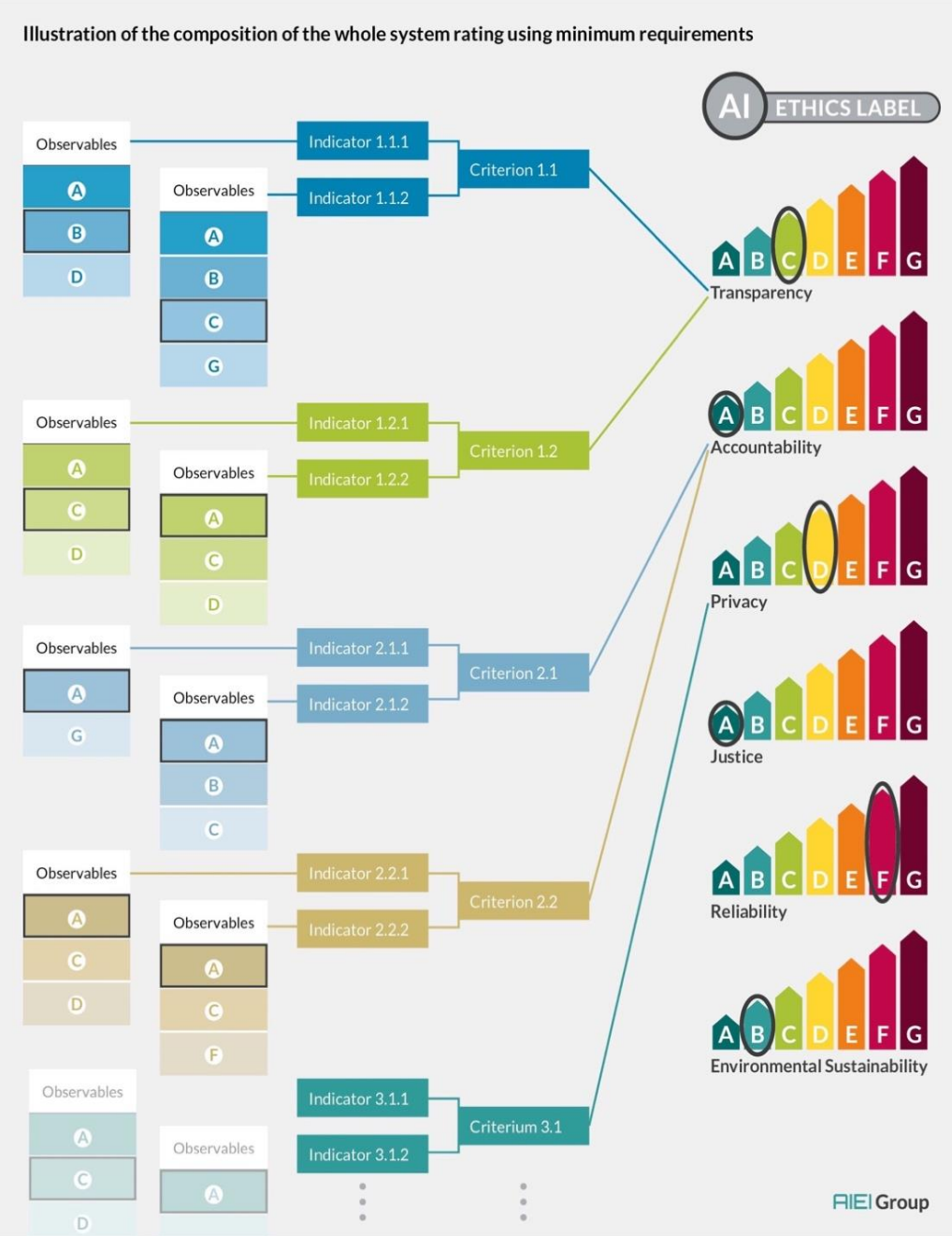


Werte formulieren ein allgemeines ethisches Anliegen, an dem sich Handeln orientieren sollte. Diese werden auf der höchsten Ebene definiert (z.B. als Wert der Transparenz). Um zu überprüfen, ob ein Algorithmus bestimmte Werte erfüllt oder verletzt, werden Kriterien festgelegt, die die Erfüllung oder Verletzung des jeweiligen Wertes definieren. Da es in der Regel nicht möglich ist, direkt zu beobachten, ob ein Kriterium erfüllt wird, braucht es die Einigung auf Indikatoren (als eine bestimmte Art von Zeichen), um dies zu überwachen. Indikatoren verbinden Kriterien mit den Observablen (Beobachtungswerten).
Die nebenstehende Grafik illustriert diese Struktur.

Ist das System-Rating nach der WKIO-Methode einmal durchgeführt kann diese Bewertung in das KI-Ethiklabel überführt werden: Das KI-Ethik Label enthält ein Rating für jeden der mit dem WKIO-Ansatz abgebildeten Werte. Das Label muss mehrere Abstufungen zulassen, um ausreichend zwischen verschiedenen Graden der Werterfüllung zu differenzieren und der Granularität der Beobachtungswerte zu entsprechen. Es sollte jedoch nicht zu viele Stufen umfassen, da dies dem Ziel, klare Orientierung über die Qualität eines Systems auf einen Blick zu bieten, zuwiderlaufen könnte. 5 bis 7 Bewertungsstufen erscheinen zweckmäßig.

Die Stufen im KI-Ethik-Label müssen für jeden Wert die detaillierteren Informationen aus dem WKIO-Modell aggregieren. Im Prinzip gibt es mehrere Möglichkeiten, dies zu erreichen. Um die gewünschten Einhaltung von Werten und deren Nachvollziehbarkeit zu gewährleisten, kommt hierbei jedoch letztlich nur die Definition von Minimalanforderungen (2) in Frage:

1. Verschiedene Observablen können einzelne metrische Werte haben, die man durch die Berechnung des Durchschnittswertes aggregiert. Schulnoten werden zum Beispiel auf diese Weise aggregiert: Ein Mathe-Test mit einer schlechten Note 4, kann durch einen anderen Test mit einer guten Note 2 kompensiert werden, was zu einer abschließenden Durchschnittsnote 3 in Mathematik führt.
2. Man kann die Mindestanforderungen an Beobachtungswerte definieren, die erforderlich sind, um eine bestimmte Systembewertung zu erreichen. Das bedeutet, dass ein durchgehender Mindeststandard in allen Zweigen (Observablen – Indikatoren – Kriterien) erreicht werden muss.



Das resultierende KI-Ethik-Label gibt auf einen Blick Auskunft über die ethisch relevanten Merkmale eines KI-Systems. Es zeigt für jeden Wert eine Bewertung an. Deren praktische Bedeutung hängt jedoch davon abhängt, in welchem Bereich das KI-System Anwendung findet. Das Szenario eines KI-Systems in einem medizinischen Kontext erfordert bspw. ein anderes Maß an Transparenz als einige industrielle Anwendungen. Gleichzeitig es ist jedoch nicht möglich, diese Anforderungen für jedes Anwendungsszenario zu berücksichtigen, genauso wenig wie es beispielsweise möglich ist, ein Strafgesetz zu schreiben, das für jeden denkbaren Fall Haftstrafen auflistet. Die Einstufung des Anwendungskontextes wiederum bestimmt den Grad der Werterfüllung, der für eine gegebene Anwendung erforderlich ist. Das Instrument hierzu ist die im nächsten Abschnitt skizzierte Risikomatrix.

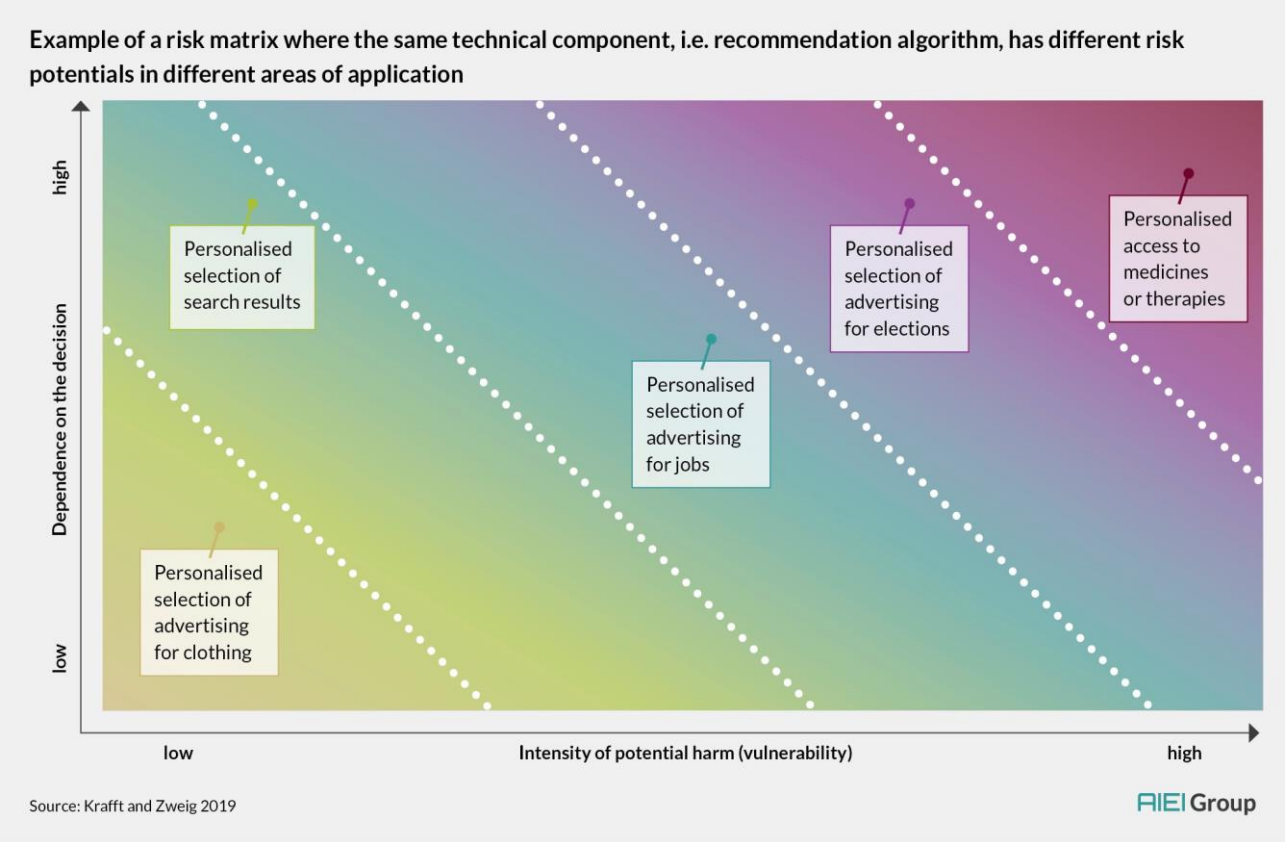
Die Risikomatrix

Nur beide Schritte zusammen - die allgemeine Beschreibung des KI-Systems mit dem Ethik-Label und eine Bewertung des Anwendungskontextes - können bestimmen, ob ein KI-System in einer gegebenen Situation ethisch vertretbar ist. Daher erfordert ein praktikabler Ansatz für den Umgang mit KI-Ethik einen komplementären Schritt zum Labelling: eine Klassifizierung der Anwendungskontexte.

Diese Klassifizierung muss auf der Grundlage des gesamten potenziellen Schadens erfolgen, den ein KI-System in seinem jeweiligen Umfeld verursachen kann. Entscheidend für die Bewertung sind also die Intensität des Schadenspotenzials des KI-Systems und die Abhängigkeit der betroffenen Person(en) von der jeweiligen Entscheidung. Hier hat sich die Verwendung einer zweidimensionalen Risikomatrix, auf der diese Faktoren die Achsen beschreiben, als geeignet erwiesen, den Klassifizierungsprozess zu vereinfachen, ohne zu sehr von der gegebenen Komplexität zu abstrahieren, in der ein KI-System arbeitet. Aus den beiden Dimensionen ergeben sich die verschiedenen Level an Transparenzpflichten und notwendigen Kontrollprozesse.

Am Beispiel von Empfehlungssystemen wie Suchmaschinen im Internet für Recherchen, auf denen dann anschließend gezielte Werbung und Kaufempfehlungen aufbauen, wird schnell deutlich, wie stark der Einfluss des Nutzungsszenarios von ADM-Systemen ist. Im folgenden Beispiel (vgl. Diagramm auf der nächsten Seite) einer Risikomatrix sind verschiedene KI-Systeme auf Basis derselben IT-Komponenten dargestellt, d.h. diese könnten mit demselben Empfehlungsalgorithmus implementiert werden. Nimmt man jedoch den Anwendungskontext hinzu, wird deutlich, dass diese ähnlichen Systeme äußerst unterschiedliche Risikopotenziale bergen.

Während in diesem Beispiel nur wenige schwerwiegende ethische Implikationen zu berücksichtigen gibt, wenn es beispielsweise um personalisierte Vorschläge für Kleidung geht, nehmen diese Implikationen mit der Intensität des potenziellen Schadens und der Abhängigkeit einer Person von der Entscheidung zu. Macht also dasselbe Empfehlungssystem nun bspw. personalisierte Vorschläge für medizinische Produkte, wird klar, dass diese Systeme unterschiedlich behandelt werden müssen.



Will eine Regulierungsbehörde die in der Risikomatrix dargestellten Nutzungsklassen verwenden, sollten unterschiedliche Anforderungen an diese KI-Systeme gestellt werden. Vorgeschlagen werden hier Klasse 0-4. In der Klasse 0, in die die meisten Systeme fallen werden, ist keine Regulierung notwendig, während in Klasse 4 keine ADM-Systeme zum Einsatz kommen sollten. Die Entscheidung, welche Art von Anwendung in welche Klasse fällt und somit bestimmten regulatorischen Anforderungen unterliegt, muss von politischen Entscheidungsträger*innen getroffen werden. Als weitere Hilfe zur Orientierung, wie dies geschehen kann, dient die folgende Beschreibung der Achsen der Risikomatrix.

X-Achse: Intensität des potenziellen Schadens / Höhe des Gesamtschadens bei Fehlurteilen
Für die x-Achse ist der kritische Aspekt das Risikopotential einer Maschine, d.h. die Intensität, mit der ein KI-System Menschen, Organisationen und der Gesellschaft schaden könnte, insbesondere durch Fehlentscheidungen. Um dieses Schadenspotenzial zu beurteilen, müssen die folgenden Aspekte berücksichtigt werden:

>> Auswirkungen auf Grundrechte, Gleichheit oder soziale Gerechtigkeit: Hat eine KI negative Auswirkungen auf die Grundrechte einer natürlichen oder juristischen Person oder sind Mechanismen der sozialen Gerechtigkeit (*z.B. Rente, Krankenversicherung*) für umfangreiche demographische Entwicklungen gefährdet oder könnten die Auswirkungen sogar katastrophal sein und zum Verlust von Menschenleben führen (*z.B. die Behandlung von Intensivpatienten*)?

>> Anzahl der betroffenen Menschen: Ist eine hohe Anzahl von Menschen betroffen (*z.B. faire Beurteilung bei einer Bewerbung, Nicht-Diskriminierung*)?
>> Auswirkungen auf die Gesellschaft: Trägt das System das Risiko, die Gesellschaft als Ganzes zu beeinträchtigen (*z.B. personalisierte Auswahl von politischen Nachrichten*), unabhängig von direkt wahrnehmbaren Schäden?

Y-Achse: Abhängigkeit von der Entscheidung / Möglichkeit der Re-Evaluierung
Die Y-Achse zeigt die Abhängigkeit der potenziell betroffenen Parteien von der algorithmischen Entscheidung und befasst sich somit mit den Optionen zur Vermeidung des potenziellen Schadens (X-Achse). Je besser die Chancen stehen, die möglichen negativen Folgen einer Entscheidung oder den durch sie verursachten Schaden zu vermeiden oder rückgängig zu machen, desto weiter unten auf der Y-Achse landet das KI-System. Die drei Hauptfaktoren, die dabei eine Rolle spielen, sind Kontrolle, Umschaltbarkeit (switchability) und Wiedergutmachung (redress).

>> Kontrolle meint, inwieweit die Entscheidungen und Handlungen eines KI-Systems zusätzlich durch sinnvolle menschliche Interaktion gefiltert werden *(z.B. der Kauf von empfohlenen Artikeln in einem Online-Shop*). Eine Anwendung mit wirkungsvoller Kontrolle impliziert einen geringeren Regelungsbedarf als eine Situation, in der Maschinen, ohne menschliche Mittler agieren (*z.B. die Notabschaltung eines Kraftwerks*).
>> Switchability oder Umschaltbarkeit ist die Möglichkeit, das KI-System gegen ein anderes auszutauschen (*z.B. durch den Wechsel des Bedieners*) oder zu vermeiden, einer algorithmischen Entscheidung ganz ausgesetzt zu sein. Ein einseitiges Abhängigkeitsverhältnis zwischen Produzenten oder Betreibern und Nutzern sowie monopolistische (auch staatliche) Strukturen führen zur Abhängigkeit von einem oder wenigen Systemen. Im schlimmsten Fall hat der Nutzer keine Möglichkeit, die Nutzung bestimmter Dienste abzulehnen, ohne sich gesellschaftlichen Folgen auszusetzen (*z.B. Gesundheitswesen, Finanzmarkt*).
>> Wiedergutmachung oder Redress betrachtet die Möglichkeit, eine automatisierte Entscheidung anzufechten oder zu korrigieren, und die Zeit, die benötigt wird, um einen solchen Antrag angemessen zu bearbeiten. Maschinelle Entscheidungen, die nicht angefochten werden können, erhöhen die Abhängigkeit der Personen von der Entscheidung. Die Behebung eines erheblichen individuellen Schadens erfordert dabei mehr Aufwand als viele Fälle von geringerem Schaden. Dieser Aspekt betrifft den besonders den *Schadensausgleich/die Haftung*.

In jedem Fall ist es nicht sinnvoll, die Intensität eines potentiellen Schadens zu bewerten, indem man lediglich die Schadenshöhe mit der Eintrittswahrscheinlichkeit multipliziert. Dies würde bedeuten, das Risiko, dass jemand im Falle eines drohenden Sturms ohne Regenschirm das Haus verlässt (hohe

Eintrittswahrscheinlichkeit, geringes Schadenspotential), mit dem Risiko eines nuklearen Unfalls (geringe Eintrittswahrscheinlichkeit, hohes Schadenspotential) gleichzusetzen. Folglich können mit zunehmendem Schadenspotential Makrorisiken entstehen, die unsere Handlungsfähigkeit überhaupt bedrohen und deshalb nicht akzeptabel sind.

Die Notwendigkeit eines multimethodischen Ansatzes

Damit ein KI-Ethik-Rahmen in der Praxis Wirkung zeigen kann, muss dieser in einem multimethodischen Ansatz Antworten auf mehrere Herausforderungen liefern:

1. **Die Werterfüllung hängt vom Anwendungsbereich und kulturellen Kontext ab**: Werte für KI-Systeme müssen durch kontextualisierte Interpretationen und ihre Anwendung auf Situationen konkretisiert werden. Wie man also Werte in der Praxis umsetzt und priorisiert, hängt bis zu einem gewissen Grad vom Anwendungsbereich und dem kulturellen Kontext ab, in dem ein KI-System operiert. Ein System, das im Justizsektor eingesetzt wird, muss bspw. notwendigerweise ein höheres Maß an Privatsphäre und Fairness aufweisen als ein System, das der Organisation der industriellen Produktion dient.

2. **Entwicklung und Implementierung eines KI-Systems beeinflussen seine Auswirkungen**: KI-Systeme sind sozio-technische Systeme. Ihre gesellschaftliche Wirkung hängt nicht nur von der Technologie (Daten und Algorithmen), sondern auch von den dem System zugrunde liegenden Zielen ab und davon, wie eine KI in eine Organisationsstruktur eingebettet ist. Die Umsetzung von Werten während des komplexen Entwicklungs- und Umsetzungsprozesses von KI-Systemen erfordert vielfältige Maßnahmen. So kann Transparenz beispielsweise von der technischen Erklärbarkeit eines KI-Systems abhängen, die in den Händen der Systementwickler liegt, aber auch eine aktive Kommunikation und Erklärung der algorithmischen Entscheidungsprozesse erfordern (Auskunfts- und Erklärungspflicht).

3. **Benutzerfreundlichkeit bedeutet unterschiedliche Dinge für verschiedene Interessengruppen**:
   Aufgrund der sozio-technischen Natur von KI-Systemen müssen Rahmenwerke für die praktische Umsetzung der KI-Ethik Werkzeuge bereitstellen, die die unterschiedlichen Rollen von Systementwicklern und Systemnutzern berücksichtigen. Solche Rahmenwerke müssen auch die externe Kontrolle über die Umsetzung dieser Maßnahmen erleichtern (Durchsetzbarkeit).

Nimmt man diese Herausforderungen als gegeben an, gibt der hier vorgestellte Ansatz die folgenden Antworten zur Lösung ethischer Anforderungen an KI-Systeme.

1. **Zur Frage der Kontextabhängigkeit**: *Kombination eines kontextunabhängigen Ethik-Ratings (WKIO) und eines kontextabhängigen Klassifikationsansatzes (Risikomatrix):*

   Jede praktische Umsetzung der KI-Ethik muss Unterschiede für die Realisierung von Werten, potenzielle Wertkonflikte und die Verschiedenheit der Anwendungskontexte berücksichtigen. Der WKIO-Ansatz ermöglicht deshalb die Bewertung ethisch relevanter Merkmale eines KI-Systems (z.B. im Hinblick auf Rechenschaftspflicht und Transparenz) unabhängig vom Anwendungskontext des Systems. In einem separaten Schritt wird eine Klassifikation verschiedener KI-Anwendungskontexte eingeführt, basierend auf dem Risiko, das sie für die betroffenen Personen und die Gesellschaft insgesamt darstellen - die Risikomatrix. Die Kombination beider Ansätze ermöglicht es, KI-Systeme ethisch zu bewerten, ohne zu entscheiden, was akzeptabel oder inakzeptabel ist. Diese Beurteilungen bleiben somit den politischen Entscheidungsträger*innen, Regulierungsbehörden und Nutzern überlassen.

2. **Zur Frage der sozio-technischen Natur von KI-Systemen**: *Bestimmung allgemeiner Prinzipien bzgl. technischer Systemanforderungen und Anforderungen der Implementierung:*

   Umsetzbare Rahmenwerke für KI-Ethik müssen den komplexen Prozess der Entwicklung und Implementierung, die sozio-technische Natur von KI-Systemen und die daraus abgeleiteten Verantwortlichkeiten von Systementwicklern und Anwendern berücksichtigen. Die Messungen (Indikatoren) und Beobachtungswerte (Observablen), die in den WKIO Ansatz einfließen, umfassen sowohl Anforderungen an das technische System (Zielgruppe: Systementwickler) als auch als Voraussetzungen für die Implementierung des Systems (Zielgruppe: Systembenutzer). Dabei umfasst das Rahmenwerk sowohl System- als auch Prozessanforderungen. Für politische Entscheidungsträger*innen, Regulierungsbehörden, Aufsichtsgremien und Überwachungsorganisationen und betroffene Verbraucher und Bürger sind beide Arten von Anforderungen relevant, um zu beurteilen, inwieweit ein automatisiertes Entscheidungssystem für ein bestimmtes Anwendungsgebiet geeignet ist oder nicht.

3. **Zur Frage der unterschiedlichen Bedürfnisse der Interessengruppen**: *Benutzerfreundlichkeit:*

   Da alle Interessengruppen unterschiedliche Verantwortungsebenen für technische und ethische Fragen haben, ist es von entscheidender Bedeutung, dass jeder Rahmen für KI-Ethik vereinfacht, aber nicht zu stark vereinfacht wird und dass er eine den Anforderungen der einzelnen Interessengruppen gerechte Orientierung bietet (auf einen Blick vs. Hintergrund).

Der Ansatz eines differenzierten KI-Ethiklabels, das sich an das bekannte Energieeffizienz-Label anlehnt beruht auf Erfahrungswerten. Labels haben sich in vielen Branchen etabliert und sich als nützlich und akzeptabel für Verbraucher, Industrie und Regulierungsbehörden erwiesen. 85 Prozent der Europäer geben an, vor dem Kauf auf das Energieeffizienz Label zu achten. Das hier vorgeschlagene KI-Ethik Label ist als Selbstverpflichtung zur Zertifizierung anwendbar, kann aber auch für eine strengere Regulierung genutzt werden und ist daher für alle beteiligten Interessengruppen praktikabel.

Die Zusammenführung

Dieser Bericht illustriert den Übergang vom "Was zum Wie", wenn es darum geht, abstrakte Werte und Prinzipien in die technische und organisatorische Praxis zu bringen. Es wurde gezeigt, wie man KI-Ethik differenziert und in einer Weise anwenden kann, die dazu beitragen kann:

- die Durchsetzung der europäischen Werte und den Schutz der Bürger*innen zu unterstützen
- Qualitätstransparenz und Vergleichbarkeit auf dem Markt zu schaffen
- den Aufwand für Unternehmen und Regulierer gering zu halten
- KI-Ethik genau dort umsetzen, wo dies erforderlich ist
- und all dies in einer Weise zu kommunizieren, die leicht zu verstehen ist.

Das abschließende Diagramm (vgl. nächste Seite) fasst die drei Hauptelemente des Modells zusammen: Links die Charakterisierung von KI-Systemen mit einem Ethik-Rating auf Grundlage des WKIO-Ansatzes (hier VCIO) und klare Vermittlung der Ergebnisse durch das KI-Ethik-Label (Mitte) sowie rechts die Klassifizierung des Anwendungskontexts in der Risikomatrix, um die regulatorischen Anforderungen für ein KI-System zu bestimmen.

Eine vertiefte Diskussion dieses Frameworks sowie weitere Aspekte wie beispielsweise die Begründung der Auswahl der Kategorien im KI-Ethik-Label findet sich in der englischen Vollversion dieses Berichts. Sie ist zum Download verfügbar unter www.ai-ethics-impact.org .

Die weitere Arbeit der AIEI Group zielt unter anderem darauf ab, das Framework auf spezifische Szenarien anzuwenden sowie das Vorgehen zur Feststellung der ethischen Charakteristika von KI-Systemen im Rahmen des VCIO-Ansatzes zu vertiefen.

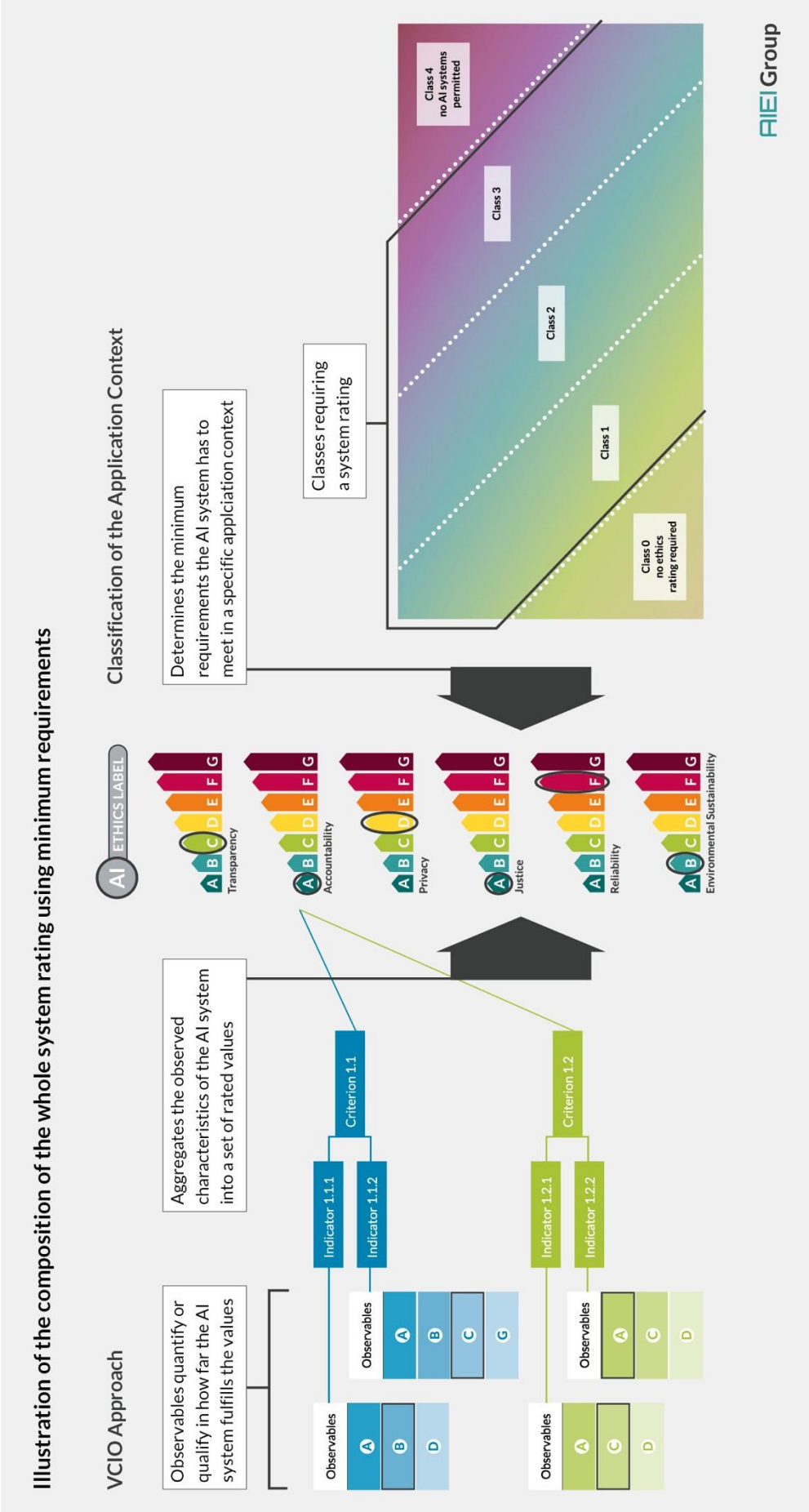


Illustration of the composition of the whole system rating using minimum requirements
VCIO Approach
Observables quantify or qualify in how far the AI system fulfills the values
Observables
Indicator 1.1.1
Indicator 1.1.2
Criterion 1.1
Indicator 1.2.1
Indicator 1.2.2
Criterion 1.2
Aggregates the observed characteristics of the AI system into a set of rated values
AI ETHICS LABEL
Transparency
Accountability
Privacy
Justice
Reliability
Environmental Sustainability
Classification of the Application Context
Determines the minimum requirements the AI system has to meet in a specific appliciation context
Classes requiring a system rating
Class 4 no AI systems permitted
Class 3
Class 2
Class 1
Class 0 no ethics rating required
AIEI Group

Über die AI Ethics Impact Group (AIEI Group)

Der VDE Verband der Elektrotechnik Elektronik Informationstechnik e.V. unter der Projektleitung von **Dr. Sebastian Hallensleben** hatte die Arbeiten an dem KI-Ethik-Label und die dahinter liegenden komplexen Überlegungen zusammen mit **Carla Hustedt** von der Bertelsmann-Stiftung initiiert. Das Projektkonsortium hat als AI Ethics Impact Group (AIEI Group) die hier in kürzerer Form vorgestellte Methode interdisziplinär erarbeitet und Expertise aus Bereichen der Informatik, Philosophie und Technologiefolgenabschätzung über Physik und Ingenieurwesen bis hin zu den Sozialwissenschaften eingebracht. Mitgewirkt haben insbesondere **Prof. Christoph Hubig** (TU Darmstadt), **Dr. Andreas Kaminski** und **Michael Herrmann** (Höchstleistungsrechenzentrum Uni Stuttgart), **PD Dr. Jessica Heesen** (Internationales Zentrum für Ethik in den Wissenschaften an der Uni Tübingen), **Dr. Thilo Hagendorf** und **Dr. Wulf Loh** (ebenfalls Uni Tübingen), **Michael Puntschuh** und **Philipp Otto** (iRights.Lab), **Andreas Hauschke** (VDE), **Prof. Rafaela Hillerbrand** (Institut für Technikfolgenabschätzung und Systemanalyse am Karlsruher KIT) mit **Paul Grünke** und **Torsten Fleischer** (ebenfalls KIT) sowie **Tobias Krafft** und **Marc Hauer** (Team von Prof. Katharina Zweig im Algorithm Accountability Lab der TU Kaiserslautern).

Das vollständige Dokument mit weiteren Aspekten der Operationalisierung ethischer Prinzipien in der Praxis kann online eingesehen werden und liegt momentan in englischer Sprache als Dokument online vor.

Kontakt:

Dr. Sebastian Hallensleben
Leiter Digitalisierung und KI

VDE e.V.
Stresemannallee 15
60596 Frankfurt am Main

Tel.: 069 6308-305
sebastian.hallensleben@vde.com

Carla Hustedt
Projektmanagement
Megatrends - Ethik der Algorithmen

Bertelsmann Stiftung
Carl-Bertelsmann-Straße 256
33311 Gütersloh

Tel.: 05241 81-81156
carla.hustedt@bertelsmann-stiftung.de

Übersetzung & redaktionelle Bearbeitung: Nora Manthey
publishing-wording.com