Bundesamt
für Sicherheit in der
Informationstechnik

Deutschland
Digital•Sicher•BSI•

Transparenz von KI-Systemen

Whitepaper

Änderungshistorie

Version	*Datum*	*Name*	*Beschreibung*
1.0	01.07.2024	Dr. Oliver Müller und Veronika Lazar	

Tabelle 1: Änderungshistorie

Bundesamt für Sicherheit in der Informationstechnik
Postfach 20 03 63
53133 Bonn
referat-t26@bsi.bund.de
Internet: https://www.bsi.bund.de

Inhalt

Vorbemerkung

Hinweis zum Sprachgebrauch: aus Gründen der Lesbarkeit wird auf die gleichzeitige Verwendung weiblicher und männlicher Sprachformen verzichtet. Alle Angaben beziehen sich auf Angehörige aller Geschlechter.

1 Einleitung

In diesem Whitepaper stellen wir eine Definition für Transparenz von informationstechnischen Systemen vor, die Künstliche Intelligenz (KI) integriert haben. Ziel dieser Publikation ist es, ein gemeinsames Verständnis über den Begriff Transparenz auszuarbeiten und die Relevanz von Transparenz für verschiedene Interessenträger sowie das BSI zu beleuchten. Daher richtet sich das Papier an alle Interessenträger von KI-Systemen und soll u.a. aufzeigen, dass unterschiedliche Interessenträger auch unterschiedliche Anforderungen an Transparenz haben können.

1.1 Motivation

KI hat sich mittlerweile sowohl im privaten als auch im beruflichen Bereich als digitales Werkzeug etabliert. Egal, ob es um die Ermittlung des persönlichen Kalorienbedarfs mittels Smartwatch, die automatische Weiterleitung von Anrufen zur Verbesserung des Kundenerlebnisses oder um die Detektion von verdächtigen Aktivitäten in Computernetzwerken geht: KI ist omnipräsent, die Liste an Beispielen für mögliche Einsatzbereiche wächst kontinuierlich. Möglich gemacht haben dies u.a. die zur Verfügung stehenden Datenmengen (engl. *big data*) zum Trainieren, Testen und Validieren der KI-Modelle, sowie mittlerweile verfügbare Hardware-Ressourcen, die entsprechende Rechenleistungen bereitstellen können. Mit den gestiegenen technischen Möglichkeiten ist auch gleichzeitig der Bedarf an KI-basierten Lösungen - insbesondere zur Effizienz- und Produktivitätssteigerung - gewachsen. Diese Nachfrage wird aktuell insbesondere durch eine zunehmende Anzahl von KI-Startups bedient. Die Folge ist eine stetig wachsende Anzahl von verfügbaren KI-Systemen, deren Technologien sich zudem rasant weiterentwickeln und deren Komplexität zunimmt.

Abstrakt betrachtet handelt es sich bei den meisten dieser Systeme um eine sogenannte Blackbox: nach außen sichtbar sind lediglich die Eingaben in das System und Ausgaben des Systems (siehe beispielsweise (Ribeiro, 2016)). Wie das System zur Ausgabe gelangt, bleibt dabei meist unklar und ist häufig nicht nachvollziehbar. Zudem ist der Wahrheitsgehalt der Ausgaben oft nicht nachprüfbar und das Verständnis über das Zustandekommen einer Systemausgabe erschwert. Die Komplexität der Systeme sowie fehlende oder mangelhafte Informationen darüber, macht sowohl eine Einschätzung per Augenschein als auch die Beurteilung der Ausgaben hinsichtlich deren Vertrauenswürdigkeit nur schwer möglich.

Bedingt durch die Techniken, die bei der Entwicklung von KI-Systemen zum Einsatz kommen, werden außerdem zusätzliche Informationen, wie beispielsweise Informationen über Trainingsdaten, relevant. So muss vor dem Einsatz von KI-Systemen die Herkunft und Qualität der Trainingsdaten eingeschätzt werden können, um beispielsweise das Risiko für Angriffe durch Datenvergiftung (engl. *poisoning attacks*), bei denen Angreifer den von einem maschinellen Lernmodell verwendeten Trainingsdatensatz manipulieren, auf das System bewerten zu können.

Zusammengefasst machen diese Faktoren die Entwicklung und den Einsatz von KI-Systemen erforderlich, die eine angemessene Nachvollziehbarkeit und Erklärbarkeit ermöglichen. Beides geht oft mit dem verwandten Kriterium der Transparenz (vgl. Abschnitt 2) einher. Transparenz ist thematisch eingebettet in das breite Feld der Vertrauenswürdigkeit von KI-Systemen. Die verschiedenen Kriterien lassen sich nicht scharf voneinander abgrenzen, beleuchten schwerpunktmäßig aber jeweils einen anderen Themenbereich. Diese Überlappung ist in Abbildung 1 dargestellt. Im vorliegenden Whitepaper geht es um das Thema Transparenz von KI-Systemen.

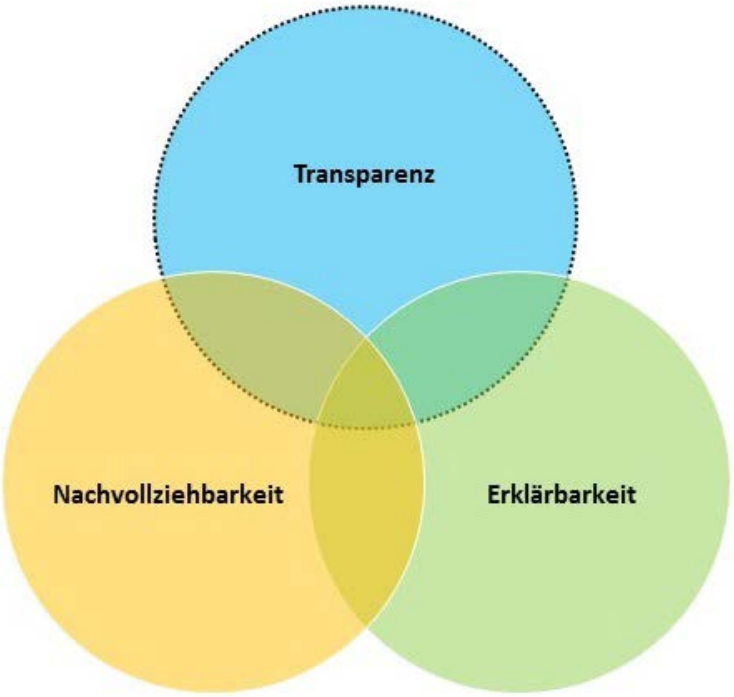

Abbildung 1: Venn-Diagramm zur Verdeutlichung des Zusammenhangs von Transparenz, Erklärbarkeit und Nachvollziehbarkeit im Kontext der Vertrauenswürdigkeit von KI-Systemen. Die verschiedenen Bereiche überschneiden sich, beleuchten aber jeweils einen eigenen Schwerpunkt.

Im Folgenden definieren wir den Begriff der Transparenz im Kontext von KI-Systemen und gehen auf die einzelnen Elemente der Definition ein. Anschließend diskutieren wir unseren Ansatz und unsere Vorgehensweise, stellen den Bezug zu den Transparenzanforderungen in der Verordnung über Künstliche Intelligenz (kurz: KI-Verordnung (KI-VO)) des Europäischen Parlaments und des Rates her und beleuchten die Chancen und Risiken transparenter KI-Systeme.

2 Definition

Transparenz von KI-Systemen ist die Bereitstellung von Informationen über den gesamten Lebenszyklus eines KI-Systems sowie über dessen Ökosystem. Transparenz forciert die Zugänglichkeit zu Informationen, die eine Einschätzung des Systems hinsichtlich unterschiedlicher Bedarfe und Ziele ermöglichen, für alle Interessenträger.

2.1 Elemente

Die obige Definition basiert auf der Vorstellung des Transparenz-Begriffs in (OECD, 2019) und (BSI, 2021a), ist konform mit den Transparenzanforderungen in der KI-VO (Details siehe Abschnitt 3.3) und repräsentiert die Position des BSI. In den folgenden Unterabschnitten werden die einzelnen Elemente der Definition genauer beschrieben und ihr Zusammenhang in Abbildung 2 grafisch veranschaulicht.

2.1.1 KI-System

Die KI-VO zur Regulierung von Künstlicher Intelligenz definiert ein KI-System als "ein maschinengestütztes System, das für einen in unterschiedlichem Grade autonomen Betrieb ausgelegt ist und das nach seiner Betriebsaufnahme anpassungsfähig sein kann und das aus den erhaltenen Eingaben für explizite oder implizite Ziele ableitet, wie Ausgaben wie etwa Vorhersagen, Inhalte, Empfehlungen oder Entscheidungen erstellt werden, die physische oder virtuelle Umgebungen beeinflussen können" (vgl. Artikel 3 KI-VO). Das BSI formuliert in seiner Definition explizit die Hardwarekomponente und definiert KI-Systeme als Software- und Hardwaresysteme, die Künstliche Intelligenz nutzen, um in der physischen oder digitalen Welt „rational“ zu handeln (BSI, 2021b). Die in diese Systeme integrierte Technologie, die als KI bezeichnet wird, umfasst verschiedene Ansätze und Techniken wie beispielsweise maschinelles Lernen, maschinelles Schließen und die Robotik. Dazu gehören u.a. Expertensysteme und Neuronale Netze, die in KI-Systemen Anwendung finden. Dies stellt keine erschöpfende Aufzählung der genutzten Techniken dar, soll jedoch die Fülle an unterschiedlichen Techniken zeigen. Eine ebensolche Bandbreite wird bei der Betrachtung der unterschiedlichen Funktionalitäten von KI-Systemen deutlich, die von einfachen bis zu hochkomplexen Aufgaben reichen. KI-Systeme können Aufgaben wie Mustererkennung, Klassifizierungen, Prognosen, Empfehlungen, natürliche Sprachverarbeitung oder auch computerbasiertes Sehen (engl. *computer vision*) abbilden und zusätzlich auf verschiedenste Art und Weise miteinander kombiniert werden. Außerdem kann KI in den Systemen je nach Anforderungen und Zielen der Anwendung auf unterschiedlicher Art und Weise implementiert sein. Sie kann einerseits als eigene Anwendung entwickelt und eingesetzt werden und die primäre Funktion des Systems darstellen, wie es beispielsweise in Chatbot-Anwendungen der Fall ist. Andererseits kann sie auch in bereits bestehende Systeme integriert sein, um zum Beispiel in Hintergrundprozessen deren Funktionalität zu erweitern und/oder die Performanz zu steigern. Betrachtet man die Art der Implementation von KI-Systemen kann auch der Automatisierungsgrad stark variieren. Während einige Systeme die Ausgaben der KI lediglich als Empfehlungen nutzen und den Menschen als finale Entscheidungsinstanz voraussetzen, setzen andere (Sub-)Systeme autonom die Entscheidungen und Klassifizierungen der KI ohne weiteres menschliches Handeln ein bzw. um. Insgesamt kann zusammengefasst werden, dass es nicht das eine KI-System gibt, sondern eine Fülle an unterschiedlichen Techniken, Funktionalitäten und Implementierungsformen betrachtet wird.

2.1.2 Ökosystem

Im vorliegenden Papier meint der Begriff Ökosystem den Kontext, in dem ein KI-System entwickelt, bereitgestellt und betrieben wird. Die Informationen bzgl. des Ökosystems eines KI-Systems gehen über das eigentliche KI-System hinaus und sollen beispielsweise auch Details über den Anbieter (z.B. Standort, Kontaktdaten) oder den Entwicklungsprozess des Systems umfassen. Der Begriff soll auch die gesamte Lieferkette des KI-Systems miteinschließen. Die Entscheidung auch Informationen über das Ökosystem eines KI-Systems in der Definition von Transparenz mit zu berücksichtigen begründet sich darin, dass eine (bedingte) Abhängigkeit zwischen dem eigentlichen KI-System und seinem Ökosystem besteht. Wird das KI-System

beispielsweise außerhalb der Europäischen Union in einem Drittstaat entwickelt und betrieben, ergeben sich entsprechende Fragen und Herausforderungen bzgl. des zugrundeliegenden IT-Sicherheits- und Datenschutzniveaus. Diese Metainformationen können eine fundierte Einschätzung der Sachlage sowie eine informierte Entscheidung durch die Interessenträger unterstützen.

2.1.3 Informationen

Informationen sind die Grundlage für das Wissen, welches die Interessenträger benötigen, um zu einer Einschätzung über das KI-System und dessen Ökosystem zu gelangen. Sie müssen offengelegt und bereitgestellt werden, damit sie dafür zur Verfügung stehen. Zudem müssen Informationen für den Wissensgewinn relevant und angemessen sein. Abschnitt 3.3 erläutert die Transparenzanforderungen in der KI-VO und zeigt auf, welche Informationen seitens der Anbieter und Betreiber von bestimmten KI-Systemen mindestens offengelegt werden müssen.

2.1.4 Lebenszyklus

Der Lebenszyklus eines KI-Systems umfasst nach (ISO/IEC 22989:2022) verschiedene Phasen, die im Folgenden im Kontext des Transparenzbegriffs kurz dargestellt werden.

2.1.4.1 Planungs- und Konzeptionsphase

Es empfiehlt sich, Transparenz schon während der Planung eines KI-Systems mitzudenken. So lassen sich zeitaufwändige Nacharbeiten in späteren Phasen des Lebenszyklus von vorne herein vermeiden. Gleichzeitig werden die beteiligten Akteure von Anfang an mitgenommen. Die Planungsphase endet mit dem Vorliegen eines konkreten Plans über das KI-System und dessen Einsatz.

2.1.4.2 Design-, Entwicklungs- und Validierungsphase

Aufbauend auf dem Plan wird das KI-System entwickelt, implementiert, getestet und validiert. Werden geplante Transparenz- und Reaktionsmaßnahmen von Anfang an mitgedacht, umgesetzt und getestet, spricht man von Transparenz durch Technologiegestaltung (engl. *transparency by design*). Dabei erlauben die Reaktionsmaßnahmen ein Nachjustieren des KI-Systems und/oder der Informationslage, wenn die Transparenz nicht im angestrebten Maße erreicht wird.

2.1.4.3 Inbetriebnahme und Anwendungsphase

Nach abgeschlossener Entwicklungs-/Validierungsphase wird das KI-System ausgerollt und in den Produktivbetrieb überführt. Während des Betriebs müssen alle für die Interessenträger relevanten Informationen bereit- sowie entsprechende Möglichkeiten für Iterationsschleifen (vgl. Abschnitt 2.1.4.4 i.V.m. Abschnitt 2.1.4.5) zur Verfügung stehen.

2.1.4.4 Kontinuierliche Evaluationsphase

Da KI-Systeme dynamisch sein können und sich die Anforderungen und/oder die Umgebung stetig verändern können, muss sich die Evaluationsphase nahtlos an den Start der Anwendungsphase anschließen. Sie läuft bei sich selbst ändernden Systemen im Sinne einer permanenten Überwachung parallel zur Anwendungsphase. Für die Evaluation unerlässlich ist die Rückmeldung der Interessenträger, z.B. bzgl. Schadens-, Not- oder Störfällen. Hierzu müssen entsprechende Reaktions- und Rückkopplungsmöglichkeiten zur Verfügung stehen (vgl. Abschnitte 2.1.4.2 und 2.1.4.3).

2.1.4.5 Systemaktualisierungen

Abhängig von den Erkenntnissen aus der Evaluationsphase (siehe Abschnitt 2.1.4.4) kann nun ggf. erneut in die Planungsphase (siehe Abschnitt 2.1.4.1) eingestiegen werden. Aktualisierungen bringen neben Fehlerbeseitigung und Leistungsverbesserung häufig auch neue Funktionen mit sich. Möglicherweise fallen auch bestehende

Funktionen weg, werden verändert oder in andere Module verlagert. Hier muss sichergestellt sein, dass bestehende Transparenz- und Reaktionsmaßnahmen durch diese Anpassungsprozesse in ihrer beabsichtigten Funktion nicht eingeschränkt oder unbrauchbar gemacht werden. Für neue Funktionen müssen entsprechend neue Maßnahmen bereitgestellt werden.

Im Falle eines erneuten Trainings der KI-Systeme (engl. *retraining*), werden weitere Maßnahmen für die Sicherstellung der Transparenz relevant. Die Maßnahmen beziehen sich dann beispielsweise auf eventuell neu verwendete Trainingsdatensätze und sind insbesondere für die Themen Diskriminierung/Bias und Geeignetheit von Relevanz.

2.1.4.6 Stilllegung

Bei der Stilllegung von Altsystemen gibt es zwei Möglichkeiten: (i) das Abschalten des Altsystems ohne Fortführung oder (ii) die Migration auf ein neues System. Hier muss jeweils transparent gemacht werden, wie mit nicht mehr benötigten bzw. migrierten Daten umgegangen wird und welche Änderungen/Konsequenzen die Stilllegung für die Interessenträger mit sich bringt. Zudem müssen die Interessenträger auch hier Reaktionsmöglichkeiten haben, um z.B. ggf. ihre Rechte als Betroffene geltend machen zu können.

2.1.5 Bedarfe und Ziele

Diese sind individuell und können in variierenden Anwendungsfällen ganz unterschiedlich ausfallen. Abgebildet werden soll mit dieser Bezeichnung, dass Transparenz nicht die Zugänglichkeit zu einer spezifischen Information ermöglichen soll, sondern dass Informationen zur Verfügung gestellt werden, die den jeweiligen Interessenträgern eine Einschätzung ermöglichen. Was konkret vom jeweiligen Interessenträger eingeschätzt wird, ist individuell und kontextabhängig. Die Bedarfe und Ziele der Interessenträger variieren je nach Anwendungsszenario. Es soll eine breite Masse an wünschenswerten Systeminformationen abgedeckt werden, die eine Einschätzung des Systems, hinsichtlich der spezifischen Bedarfe der Interessenträger, ermöglichen.

2.1.6 Interessenträger

Der Begriff Interessenträger meint alle Beteiligten, die entweder mittelbar (d.h. nicht direkt, aber z.B. durch Auswirkungen) oder unmittelbar (d.h. direkt, z.B. durch Anwendung) von einem KI-System betroffen sind oder die auf das System einwirken (z.B. Entwickler). Dies können einzelne Personen oder auch Gruppen von Personen sein. Ein Interessenträger muss dabei keine „aktive“ Rolle einnehmen. Eine Übersicht über mögliche verschiedene Interessenträger sowie deren Bezug zu KI-Systemen ist in Tabelle 2 dargestellt. Während Verbraucher und Anwender ein KI-System i.d.R. lediglich verwenden, ist es möglich, dass Experten, Entwickler und Unternehmen/Organisationen ein KI-System zusätzlich bereitstellen. Mittelbar Betroffene/Dritte stellen weder ein KI-System bereit, noch verwenden sie es. Dennoch können sie von den Auswirkungen betroffen sein und so zu (passiven) Interessenträgern werden. Die vorgestellte Auflistung der verschiedenen Interessenträger erhebt keinen Anspruch auf Vollständigkeit und kann beliebig verfeinert werden. Die gewählte Darstellung ist jedoch ausreichend um zu zeigen, dass unterschiedliche Interessenlagen bzgl. eines KI-Systems vorhanden sein können, die sich u.a. in unterschiedlichen Anforderungen an die Transparenz - wie z.B. die Art oder den Detailgrad der bereitgestellten Informationen - eines KI-Systems spiegeln können. Daher müssen die unterschiedlichen Interessenträger bei der Definition des Begriffes Transparenz zwingend berücksichtigt werden.

Interessenträger	*verwenden*	*bereitstellen*	*Beispiele für Anforderungen an Transparenz*
Verbraucher	+	-	Serverstandort in Hinblick auf den Datenschutz
Anwender	+	-	Gebrauchsanweisung zur Vermeidung von Anwendungsfehlern
Experten	+/-	+/-	Funktionsweise des zugrundeliegenden Modells
Entwickler	+	+	Technische Dokumentation zur Identifikation von Schnittstellen
Unternehmen/Organisationen	+	+	Lizenzbedingungen zur Vermeidung von strafrechtlichen Konsequenzen
Mittelbar Betroffene/Dritte	-	-	Ansprechpartner im Schadensfall

Tabelle 2: Beispielhafte Darstellung möglicher Interessenträger von KI-Systemen sowie deren unterschiedlichen Anforderungen an Transparenz aufgrund verschiedener Interessenlagen. Verwendete Zeichen: „+“ (trifft zu), „-“ (trifft nicht zu).

3 Diskussion

3.1 Ansatz und Vorgehen

Das Vorliegen bereits veröffentlichter Definitionen des Transparenz-Begriffs wirft die Frage nach der Notwendigkeit einer weiteren Definition auf. Die schiere Breite zum Thema Transparenz auf dem Definitionsmarkt spiegelt jedoch letztendlich die unterschiedlichen Anforderungen an Transparenz in Abhängigkeit der Interessenträger und des Einsatzgebietes wider. Um eine Grundlage für die weiteren Arbeiten des BSI mit Fokus auf breite Interessenträgergruppen und generischen Einsatzgebiete zu schaffen, wurde von der Nutzung bereits vorhandener Definitionen abgesehen.

Zudem ist die Geschwindigkeit, mit der sich Technologien im KI-Bereich weiterentwickeln, enorm. Dies birgt die Gefahr, dass einmal festgeschriebene Definitionen ihre Gültigkeit verlieren können, insbesondere dann, wenn sie zu spezifisch sind. Um mit dem technischen Fortschritt schritthalten zu können und um ein ständiges Erneuern und Anpassen der Definition an den aktuellen Stand der Technik zu vermeiden, ist die in diesem Whitepaper vorgestellte Definition von Transparenz so technologieneutral und zukunftstauglich wie möglich. Zum einen soll sie gut verständlich sein, alle relevanten Aspekte von Transparenz abdecken und gleichzeitig offen genug gehalten sein, um die individuelle Interpretation in Abhängigkeit des jeweiligen Interessenträgers und der jeweils eingesetzten KI-Technologie zu ermöglichen. Zum anderen soll sie als generische Grundlage für zukünftige Arbeiten des BSI in diesem Bereich dienen.

Des Weiteren wurde bei der Definition ein holistischer Ansatz verfolgt. Dieser ist in Abbildung 2 dargestellt: Transparenz beinhaltet sowohl die Bereitstellung von Informationen über das KI-System selbst als auch über dessen Ökosystem, wie z.B. die Lieferkette des KI-Systems oder Details über den Anbieter. Die Gewichtung der bereitgestellten Informationen obliegt dem jeweiligen Interessenträger.

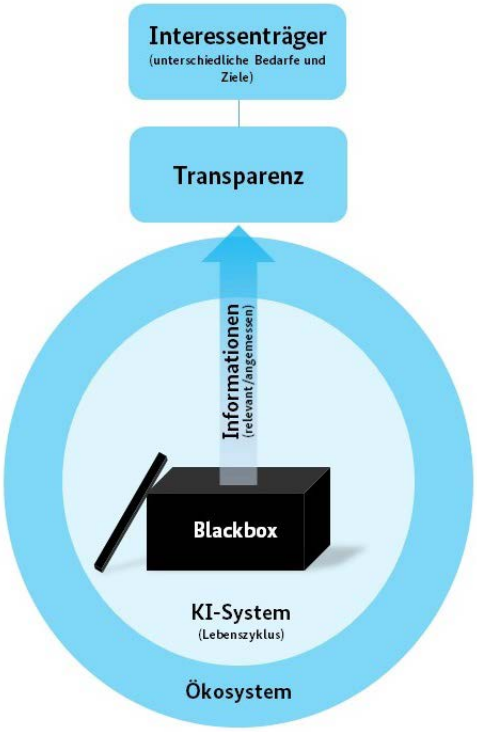

Abbildung 2: Schematische Darstellung des holistischen Transparenzansatzes unter Verwendung der einzelnen Elemente aus der Definition: neben für die Interessenträger relevanten und angemessenen Informationen zum KI-System selbst werden auch Informationen über dessen Ökosystem bereitgestellt/offengelegt. Dies fördert eine valide Einschätzung der Geeignetheit und Angemessenheit des KI-Systems durch die Interessenträger.

3.2 Ziel von Transparenz

Durch das Forcieren von Transparenz von KI-Systemen soll die Autonomie der Interessenträger gestärkt und diese dazu befähigt werden, selbst zu entscheiden, ob die Verwendung, Veränderung oder Bereitstellung eines KI-Systems für sie angemessen und vertretbar ist. Dabei genügt es nicht, sich auf die Darstellung der Fähigkeiten des KI-Systems zu beschränken. Auch die Limitierungen des Systems müssen in Augenschein genommen und transparent gemacht werden. Nur so kann eine ganzheitliche Einschätzung durch die Interessenträger vorgenommen werden, z.B. ob ein KI-System für einen bestimmten Zweck geeignet ist oder nicht.

Im Bereich des digitalen Verbraucherschutzes soll dies dazu beitragen, dass Verbraucher trotz einer zunehmenden Digitalisierung sichere und vertrauenswürdige KI-Systeme erkennen und für sich nutzen können. Unternehmen und Organisationen sollen darüber hinaus in die Lage versetzt werden, eigene KI-Systeme transparent zu entwickeln und zu betreiben. Auch von Auswirkungen betroffene Dritte sollen durch transparente Informationen aus dem Ökosystem eines KI-Systems erkennen können, wie sie im Schadensfall ihre Rechte geltend machen können. Somit dient die Transparenz von KI-Systemen der Bemächtigung von Interessenträgern (engl. *empowerment*).

3.3 Transparenzanforderungen in der KI-VO

Das weltweit erste umfassende gesetzliche Regelwerk für KI ist die KI-VO. Sie wurde am 21. Mai 2024 vom Rat der Europäischen Union (EU) verabschiedet und regelt den Einsatz von KI in der EU. Dieses Whitepaper bezieht sich auf die zum Zeitpunkt des Schreibens aktuelle Fassung vom 13. Juni 2024. Die KI-VO führt Transparenz als Kernanforderung an, um einen so genannten ethischen und verantwortungsvollen Umgang mit Daten zu gewährleisten. An KI-Systeme, die in gewisse Risikostufen fallen, werden demnach Anforderungen hinsichtlich einem angemessenen Maß an Transparenz und hinreichend transparentem Betrieb der Systeme gestellt.

In der KI-VO werden u.a. harmonisierte Transparenzvorschriften für bestimmte KI-Systeme festgelegt (siehe KI-VO Artikel 1 Absatz 2 Buchstabe d). In Artikel 13 der KI-VO ist festgelegt, dass Hochrisiko-KI-Systeme (z.B. im Bereich der Biometrie oder der kritischen Infrastruktur (vgl. KI-VO Anhang III)) so transparent sein müssen, dass Anbieter und Betreiber ihre ebenfalls in der KI-VO festgelegten einschlägigen Pflichten erfüllen können. Hierzu legt die KI-VO auch fest, welche Informationen in der Betriebsanleitung mindestens enthalten sein müssen. Die KI-VO legt auch fest, dass Anbieter und Betreiber von KI-Systemen, die für die direkte Interaktion mit natürlichen Personen bestimmt sind (z.B. Chatbot-Anwendungen), die betroffenen natürlichen Personen ebenfalls darüber informieren müssen, dass es sich um ein KI-System handelt oder dass die System-Ausgabe KI-generiert ist (vgl. KI-VO Artikel 50). Durch diese Offenlegung und Bereitstellung von Informationen für verschiedene Interessenträger soll Transparenz hinsichtlich des jeweiligen KI-Systems geschaffen werden. Dabei stellt sowohl diese spezifische Art von Systemen als auch die reine Kenntlichmachung des Einsatzes von KI-Systemen nur eine Teilmenge dessen dar, was unter der die in diesem Whitepaper vorgestellte Definition von Transparenz verstanden wird.

Zur Durchführung der KI-VO und für die praktische Umsetzung der Transparenzpflichten gemäß Artikel 50 erarbeitet die Kommission Leitlinien (siehe Artikel 96 Absatz 1 Buchstabe d). Das Vorgehen bei Sanktionen, die bei Verstößen gegen die KI-VO wirksam werden, ist in Artikel 99 festgelegt. Verstöße gegen die in Artikel 50 festgelegten Transparenzpflichten für Anbieter und Betreiber sind noch einmal explizit erwähnt (siehe KI-VO Artikel 99 Absatz 4 Buchstabe g), was die Relevanz der Thematik in der KI-VO unterstreicht.

Im Zuge der Bewertung und Überprüfung der KI-VO bewertet die Kommission alle vier Jahre Änderungen der Liste der KI-Systeme, die zusätzliche Transparenzmaßnahmen erfordern (gemäß Artikel 50 KI-VO). Zudem soll „eine partizipative Methode für die Bewertung der Risikoniveaus anhand der in den jeweiligen Artikeln genannten Kriterien und für die Einbeziehung neuer Systeme“ in diese Liste erstellt werden (KI-VO Artikel 112 Absatz 11 Buchstabe c).

In Anhang XII der KI-VO geht es um „Transparenzinformationen gemäß Artikel 53 Absatz 1 Buchstabe b [für die Erstellung] (...) technische[r] Dokumentation[en] für Anbieter von KI-Modellen mit allgemeinem Verwendungszweck für nachgelagerte Anbieter, die das Modell in ihr KI-System integrieren“. Dabei bezieht sich

Absatz 1 auf die Informationen über das Modell, die mindestens in der Dokumentation enthalten sein müssen. Während Absatz 1 sich auf das KI-System selbst bezieht, geht Absatz 2 auch über das eigentliche KI-System hinaus und fordert z.B. Informationen über Bestandteile des Entwicklungsprozesses des Modells. Auch an dieser Stelle sind die in der KI-VO festgelegten Transparenzpflichten mit der in diesem Whitepaper vorgestellten Definition von Transparenz konform, da Informationen über das Ökosystem mit einbezogen werden.

Abgesehen von den konkreten Transparenzanforderungen wird in der KI-VO der Zweck der KI-VO in Artikel 1 Absatz 1 niedergeschrieben. Neben der Verbesserung des Funktionierens des Binnenmarkts soll die KI-VO „die Einführung einer auf den Menschen ausgerichteten und vertrauenswürdigen“ KI fördern. Gleichzeitig soll ein hohes Schutzniveau vor schädlichen Auswirkungen von KI-Systemen in der EU gewährleistet werden. Die in diesem Whitepaper vorgestellte Definition von Transparenz darf diesen Zielen nicht entgegenstehen. Die Bereitstellung von Informationen über das KI-System und dessen Ökosystem soll dazu beitragen, die Vertrauenswürdigkeit eines KI-Systems zu erhöhen. Zudem kann Transparenz das Zusammenspiel verschiedener Akteure/Interessenträger entlang der Lieferkette vereinfachen, was wiederum dem Funktionieren des Binnenmarktes dienlich sein kann. Gleichzeitig darf Transparenz nicht dazu führen, dass KI-Systeme durch die Offenlegung bestimmter (z.B. sicherheitsrelevanter) Informationen von Angreifern missbraucht werden können (mehr dazu in Abschnitt 3.5). Die KI-VO legt außerdem fest, dass Vorfälle oder Fehlfunktionen im Zusammenhang mit KI nicht dazu führen dürfen, dass die Gesundheit oder die Sicherheit von Personen oder Eigentum durch ein KI-System gefährdet wird. Weitere Ziele der KI-VO sind die Förderung der „in der Charta verankerten Grundrechte, einschließlich Demokratie, Rechtsstaatlichkeit und Umweltschutz“. Hier kann Transparenz dazu beitragen, dass die Interessenträger zu einer Einschätzung über die Geeignetheit eines KI-Systems gelangen können. Diese Bemächtigung versetzt die Interessenträger in die Lage selbst zu entscheiden, ob ein KI-System in Hinblick auf die genannten Punkte eingesetzt werden kann/soll oder nicht. Durch das Offenlegen von Verantwortlichkeiten kann Transparenz im Schadensfall dazu beitragen, den Schaden einzudämmen und möglichen Folgen vorzubeugen bzw. diese zu minimieren. Transparenz kann zudem ein Innovationstreiber sein. Das Wissen über Limitierungen von KI-Systemen kann in der Entwicklung von Anwendungen/Produkten münden, die diese Limitierungen nicht mehr aufweisen. Beispielsweise war die Interaktion mit den ersten Chatbots anfänglich nur über Textsprache möglich: über eine Tastatur konnten Eingaben getätigt werden, auf die der Chatbot dann auf dem Bildschirm in Textsprache reagiert hat. Mittlerweile kommen in verschiedensten Bereichen (z.B. bei der telefonischen Kundenbetreuung im Versicherungsbereich) Chatbots zum Einsatz, bei denen eine Interaktion mittels Spracheingabe und Sprachausgabe möglich ist.

Zusammenfassend lässt sich sagen, dass Transparenz in der KI-VO Beachtung findet und erste Anforderungen formuliert werden. Zugleich wird der Transparenzbegriff in der KI-VO sehr breit gefasst. Die in diesem Whitepaper vorgestellte Definition von Transparenz widerspricht diesen in der KI-VO festgelegten Transparenzforderungen nicht, soll jedoch eine besser greifende Formulierung und Umgrenzung des Begriffs Transparenz bieten.

3.4 Chancen durch Transparenz

Der Einsatz transparenter KI-Systeme kann die Nachvollziehbarkeit von Entscheidungen und die Beurteilung der Angemessenheit von Systemen fördern. Transparenz kann zudem zum Schutz vor Missbrauch beitragen, indem sie ermöglicht, potenzielle Risiken und unerwünschte Auswirkungen frühzeitig zu erkennen. Um angemessen auf Problematiken reagieren zu können ist es z.B. wichtig zu wissen, ob die Ausgabe eines KI-Systems frei von Diskriminierung ist oder ob sie gegen Lizenzbedingungen verstößt. Auch im Hinblick auf den Verbraucherschutz kann Transparenz als Unterstützungswerkzeug fungieren. Transparenz stellt die Grundlage für eine korrekte Einschätzung der Angemessenheit des verwendeten Systems dar. Um eine solche Einschätzung überhaupt vornehmen zu können, müssen Informationen über das System zugänglich sein. Eine valide Einschätzung der Angemessenheit des KI-Systems bildet die Basis für positiv verlaufende Vertrauens- und Akzeptanzprozesse. Erste Publikationen zeigen beispielsweise höhere Downloadzahlen transparenter KI-Modelle, was als Indiz für eine bessere Akzeptanz dieser Systeme bei Entwicklern sein könnte (Liang, 2024). Fehlende Transparenz erschwert die valide Einschätzung der Angemessenheit eines Systems, und somit die Einschätzung dessen Vertrauenswürdigkeit. Letzteres ist eine Grundvoraussetzung für die Entstehung und Aufrechterhaltung einer

positiven Vertrauensbeziehung zu dem System und den damit verbundenen Ausgaben. Den Nutzern kann Transparenz zusätzlich ermöglichen, Rechte leichter geltend zu machen, indem die Forderung nach Transparenz mit klareren Definitionen rechtlicher Verantwortlichkeiten und einer Identifizierung von Verantwortlichen für die Verwendung von KI-Systemen einhergeht. Die aufgeführten Aspekte Nachvollziehbarkeit, Missbrauchsschutz, Akzeptanz und Vertrauenswürdigkeit sowie rechtliche Verantwortlichkeit zeigen die Relevanz von Transparenz beim Einsatz von KI-Systemen. Diese Relevanz spiegelt sich auch in regulatorischen und rechtlichen Vorgaben (vgl. Abschnitt 3.3).

Transparenz kann zum einen zur Sicherheit von KI-Systemen beitragen und andererseits eine sichere Nutzung von KI-Anwendungen fördern. So kann Transparenz die Identifikation möglicher Probleme und Schwachstellen ermöglichen, unerwünschtes Systemverhalten sichtbar machen und zur Problemerkennung sowie Prävention von missbräuchlichem Einsatz beitragen. Transparenz liefert im Kontext von IT-Sicherheit zudem die Basis für die Offenlegung und Bewertung von Risiken, die mit dem Einsatz und der Nutzung des Systems einhergehen. Auch die Kenntlichmachung von Rollen und Verantwortlichkeiten bei Schadensfällen und unerwünschten Ereignissen im Zuge von Transparenzanforderungen kann Interessenträger dabei unterstützen, fehlerhaftes Systemverhalten zu detektieren, Reaktionszeiten zu verkürzen und so mögliche Folgeschäden einzudämmen.

Wird Transparenz schon in frühen Phasen des Lebenszyklus eines KI-Systems gelebt, so können innerhalb des Entwicklerteams von Beginn an Inkonsistenzen vermieden, Fehlerquellen minimiert und Einarbeitungsphasen verkürzt werden. Dabei werden KI-Systeme sowohl entwickelt als auch zunehmend als Entwicklungstools - z.B. beim KI-unterstützten Programmieren zur automatischen Generierung von Programmcode - für neue Systeme eingesetzt. In den frühen Entwicklungsphasen ist es aus Entwicklersicht zudem wichtig zu wissen, woher die Trainings-/Test-/Validierungsdaten stammen, wie diese beschaffen sind und ob diese frei von Bias - z.B. zur Vermeidung von Diskriminierung - sind. Diese Informationen sind wichtig, um die Daten vor dem Trainieren/Testen/Validieren korrekt aufbereiten (engl. *preprocessing*) zu können. Aktuelle Entwicklungstrends zeigen außerdem, dass häufig bereits existierende Modelle verwendet werden, was das Vorhandensein und die Zugänglichkeit aller sicherheitskritischen Informationen zu den Bestandsmodellen besonders relevant macht. Fehlen diese Informationen besteht die Gefahr, die Sicherheitsrisiken der Grundmodelle in die eigenen Produkte zu implementieren. Beide Systeme stehen dann in einer Abhängigkeit und Intransparenzen des Basis-Systems werden so auf das aufgesetzte System übertragen. Die Vererbung der oben beschriebenen Sicherheitsrisiken aus verschiedenen Bereichen führt in den genannten Beispielen zu einem erhöhten Gesamtrisiko, was die Brisanz von Transparenz für sicherheitsrelevante Aspekte bei der Nutzung von KI-Systemen noch einmal betont.

Zusätzlich kann Transparenz, wie bereits im vorangegangenen Abschnitt erörtert, zu einer besseren Befähigung der Nutzer beitragen, KI-Systeme einzuschätzen. Eine korrekte Einschätzung der Anwendung, passender Einsatzszenarien und möglicher Problematiken und Sicherheitsrisiken, kann eine sichere Nutzung durch die Anwender fördern.

3.5 Gefahren durch Transparenz

Bisher wurden überwiegend die positiven Seiten von Transparenz dargestellt. Eine Erhöhung/Verbesserung der Transparenz von KI-Systemen kann aber auch ungewollte negative Effekte mit sich bringen. So können beispielsweise durch das Bereitstellen von Informationen zur Funktionsweise oder Architektur eines KI-Systems neue Angriffsvektoren offengelegt werden, die Angreifende für den Missbrauch oder die Kompromittierung des Systems ausnutzen können.

Informationen über Limitierungen oder ausgeschlossene Anwendungsbereiche eines KI-Systems könnten von Angreifenden ebenfalls bewusst ausgenutzt werden, z.B. zur bewussten Erzeugung von fehlerhaftem Verhalten oder destruktiven Ausgaben.

Umgekehrt können Angreifende das Vertrauen, das Transparenz schaffen soll, auch dazu missbrauchen, um bewusst fehlerhafte Informationen vermeintlich transparent bereitzustellen. So können beispielsweise in Wahrheit sicherheitskritische Anwendungen als unkritisch dargestellt werden. Zusätzlich können auch intransparente Systeme als transparent gekennzeichnet werden. Eine solche Pseudotransparenz kann zu Vermarktungszwecken des eigenen Produktes genutzt werden und eine falsche Einschätzung des Systems

seitens der Verbraucher zur Folge haben, wenn keine Überprüfung des Labels Transparenz erfolgt. Daher müssen zukünftig auch die Fragen nach der Vertrauenswürdigkeit der offengelegten/bereitgestellten Informationen beantwortet werden. Ein offizielles Transparenz-Gütesiegel sowie überprüfbare Transparenzkriterien könnten hier Abhilfe schaffen.

Transparenz von KI-Systemen ist also ein zweischneidiges Schwert und daher mit Vorsicht zu verwenden. Die Ziele und Probleme sind zum Teil gegenläufig und können nicht simultan gelöst werden. Hilfreich dabei kann die Beantwortung der Leitfragen "Welche Informationen benötigt ein Interessenträger zur Entscheidungsfindung?" und "Welche Informationen sind nicht von Relevanz?" sein. Ähnlich wie in der EU Datenschutzgrundverordnung empfiehlt sich auch hier die Anwendung des Prinzips der Datensparsamkeit, welches für jeden individuellen Anwendungsfall separat beantwortet wird: es sollten so viele Informationen wie notwendig, aber nicht mehr als unbedingt erforderlich preisgegeben werden. Dieses "need-to-know"-Prinzip gilt besonders bei sicherheitskritischen Informationen. Ziel sollte ein angemessenes Level an Transparenz sein, welches ausreichend ist, und gleichzeitig Aspekte wie z.B. Sicherheit nicht zu sehr benachteiligt.

4 Schlussfolgerungen

Durch die Blackbox-Eigenschaften vieler KI-Systeme werden Daten und Informationen auf für die Nutzer nicht transparente Art und Weise verarbeitet und im Anschluss eine unüberprüfbare Entscheidung ausgegeben. Fehlendes Wissen über das KI-System geht mit fehlender Nachvollziehbarkeit und Überprüfbarkeit der System-Ausgaben einher. Eine Einschätzung, ob korrekte und angemessene System-Ausgaben erfolgen, gestaltet sich schwierig. Genauso wenig können Fragen zur Verantwortlichkeit, Haftung oder Fairness beantwortet werden, wenn es an Informationen über das System und dessen Ökosystem mangelt. Schließlich können intransparente KI-Systeme zu Vertrauensverlust und einer Ablehnung des Systems führen. Die Implementierung von KI-Komponenten in bestehende Systeme und die Kombination verschiedener Systeme kann die Komplexität zudem weiter steigern und erschwert die Zugänglichkeit zu relevanten Informationen zusätzlich. Die Probleme, die durch mangelnde Systemeinsicht und fehlende Informationen über das System entstehen sind mannigfaltig und stellen eine große Herausforderung dar. Transparenz setzt an dieser Problematik der mangelnden Informationen an und will KI-Systeme durch Erhöhung der Zugänglichkeit zu Systeminformationen nachvollziehbarer machen und eine valide Einschätzung der Systeme ermöglichen. Aus diesen Gründen spielt Transparenz für alle Interessenträger eines KI-Systems eine entscheidende Rolle. Die Herausforderung besteht darin, alle Interessenträger mit ihren individuellen und unterschiedlichen Anforderungen an Transparenz gleichsam zu bedienen.

Das Gesamtvorhaben und zukünftige Arbeiten im Bereich Transparenz von KI-Systemen richten sich an alle Interessenträger des BSI. Die Relevanz der Thematik für die Gesellschaft als Nutzer der Systeme zeigt sich in der erwarteten höheren Nachvollziehbarkeit, besserem Missbrauchsschutz, valideren Akzeptanz- und Vertrauenswürdigkeitsprozessen sowie einer verbindlicheren rechtlichen Verantwortlichkeit. Die Transparenzmaßnahmen sollen direkt zur Befähigung der Endanwender beitragen, indem deren Vertrauen sowie Autonomie bezüglich Wahl und Nutzung von KI-Systemen gestärkt wird. Insgesamt zielt diese Befähigung der Endanwender auf eine Demokratisierung des Einsatzes von KI-Systemen ab. Zudem soll die Arbeit im Bereich Transparenz und die Ableitung konkreter Kriterien und Maßnahmen zum übergeordneten Ziel des vertrauenswürdigen Einsatzes von KI-Systemen beitragen. Für Unternehmen im Bereich der Entwicklung von KI-Systemen soll die Relevanz der Thematik und die Beachtung der Maßnahmen bei Entwicklung und Betrieb der KI-Systeme forciert werden. Für die Interessenträger aus dem wirtschaftlichen Umfeld, die fremde KI-Systeme in ihren Organisationen nutzen oder in ihre Systeme und Produkte implementieren wollen, sollen Richtlinien und Positionen als Orientierungshilfen zur Verfügung gestellt werden. Diese Orientierungshilfen sollen den Unternehmen die Identifikation von geeigneten, sicheren und performanten Systemen erleichtern. Auch für Organe der öffentlichen Hand, die KI-Systeme nutzen wollen, sollen diese Arbeiten als Orientierungshilfen dienen. Zusätzlich zur eigenen Nutzung stellen die täglich neuen sicherheitlich relevanten Erkenntnisse zu KI-Systemen, die es zu adressieren gilt, die Interessenträger der öffentlichen Hand und Verwaltung vor die Herausforderung der Sicherstellung einer fachlich qualifizierten und adäquat aufgestellten Personaldecke. Diese und zukünftige Arbeiten auf dem Gebiet der Transparenz können dazu eingesetzt werden, permanente und adäquate (Nach)Schulungen des Personals zu erleichtern und zu forcieren. Zudem kann die durch diese und zukünftige Arbeit erhoffte Etablierung von Transparenzkriterien die Entwicklung von aussagekräftigen und verlässlichen Gütesiegeln durch Organe der öffentlichen Hand erleichtern. Im Hinblick auf eine zu erwartende weiter zunehmende Verbreitung und flächendeckendes Ausrollen von KI-Systemen in viele Lebensbereiche nimmt die gesamtgesellschaftliche Relevanz stetig zu.

Um zukünftig kompetente und valide Einschätzungen dieser Systeme vornehmen zu können, ist die Etablierung von Transparenzkriterien unabdingbar. Für Anbieter und Betreiber bestimmter KI-Systeme - wie z.B. von KI-Systemen mit allgemeinem Verwendungszweck oder von Emotionserkennungssystemen - sind in der KI-VO bereits Transparenzpflichten definiert (vgl. Artikel 50 KI-VO). Diese sind eine der Voraussetzungen dafür, dass diese Systeme in der Europäischen Union vertrieben und verwendet werden dürfen. Transparenzkriterien können die Autonomie der Interessenträger eines KI-Systems stärken, indem informierte Entscheidungen möglich gemacht werden. Daher kann und sollte Transparenz von Anfang an mitgedacht werden (Transparenz durch Technologiegestaltung, engl. *transparency by design*).

Literaturverzeichnis

BSI, Bundesamt für Sicherheit in der Informationstechnik. 2021a. AI Cloud Service Complience Criteria Catalogue (AIC4). 2021.

BSI, Bundesamt für Sicherheit in der Informationstechnik. 2021b. Bundesamt für Sicherheit in der Informationstechnik, Sicherer, robuster und nachvollziehbarer Einsatz von KI - Probleme, Maßnahmen und Handlungsbedarfe. 2021.

ISO/IEC 22989:2022. Information technology - Artificial intelligence - Artificial intelligence concepts and terminology.

Liang, W., Rajani, N., Yang, X., Ozoani, E., Wu, E., Chen, Y., Smith, D. S., & Zou, J. 2024. What's documented in AI? Systematic Analysis of 32K AI Model Cards. 2024. http://arxiv.org/abs/2402.05160.

OECD. 2019. Recommendation of the Council on Artificial Intelligence. *Recommendation of the Council on Artificial Intelligence.* 2019.

Ribeiro, M. T., Singh, S., & Guestrin, C. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. s.l. : Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135–1144. https://doi.org/10.11, 2016.