Universität Zürich[UZH]

Universität Basel

Juristische Fakultät

WHITE PAPER

Transparenz

Florent Thouvenin
Stephanie Volz

Juli 2024

CENTER FOR INFORMATION TECHNOLOGY SOCIETY AND LAW — ITSL

e-PIAF
electronic Public Institutions and Administrations Research Forum

Dieses White Paper wurde im Projekt «**Nachvollziehbare Algorithmen: ein Rechtsrahmen für den Einsatz von Künstlicher Intelligenz**» entwickelt, das vom Center for Information Technology, Society, and Law (ITSL) der Universität Zürich und von der Forschungsstelle electronic Public Institutions and Administrations Research Forum (e-PIAF) der Universität Basel durchgeführt und von der Stiftung Mercator finanziell unterstützt wird. Dieses White Paper ist Teil einer Reihe von White Papers, die sich mit den zentralen Herausforderungen befassen, die mit dem Einsatz von Künstlicher Intelligenz (KI) in Unternehmen und in der Verwaltung verbunden sind.

Die White Papers und weitere Informationen zum Projekt finden sich auf **www.itsl.uzh.ch** und **www.ius.unibas.ch/e-piaf**.

Folgende White Papers sind bislang erschienen:

- **Manipulation**
- **Diskriminierung**
- **Datenschutz**
- **Transparenz**
- **Transparenz durch Begründung von Verfügungen**
- **Transparenz durch öffentliches Verzeichnis**

Das Kernprojektteam besteht aus folgenden Personen:

Prof. Dr. Florent Thouvenin, Professor für Informations- und Kommunikationsrecht an der Universität Zürich, Vorsitzender des Lenkungsausschusses des ITSL

Prof. Dr. Nadja Braun Binder, MBA, Professorin für öffentliches Recht an der Universität Basel

Dr. Stephanie Volz, wissenschaftliche Geschäftsführerin des ITSL und Lehrbeauftragte an der Universität Zürich

Liliane Obrecht, MLaw, Wissenschaftliche Mitarbeiterin und Doktorandin an der Juristischen Fakultät der Universität Basel

Beim Einsatz von Künstlicher Intelligenz (KI) wird häufig Transparenz gefordert. Zum einen soll für Menschen erkennbar sein, dass sie nicht mit einem anderen Menschen, sondern mit einem KI-System interagieren, oder dass ein Text, ein Bild, ein Video oder ein anderer Inhalt von KI generiert wurde («Erkennbarkeit»). Zum andern sollen Menschen verstehen können, wie ein KI-System zu einer bestimmten Vorhersage, Empfehlung oder Entscheidung gelangt ist oder wie es einen bestimmten Inhalt erstellt hat («Erklärbarkeit und Nachvollziehbarkeit»). Transparenz ist auch bei der Rechtsdurchsetzung relevant, weil Kläger die Tatsachen nachweisen müssen, auf die sie ihre Klage stützen. Das geltende Recht gewährleistet die Transparenz beim Einsatz von KI-Systemen teilweise, aber nicht umfassend. Neben einer Anwendung der bestehenden Regeln des Datenschutzgesetzes (DSG) und der Zivilprozessordnung (ZPO) sind punktuelle Ergänzungen des geltenden Rechts sinnvoll. Im Vordergrund stehen eine ausdrückliche Regelung der Erkennbarkeit bei einer Interaktion mit einem KI-System (bspw. ein Chatbot) und eine Ausdehnung der bestehenden Informationspflicht bei vollautomatisierten Einzelentscheidungen auf weitgehend automatisierte Einzelentscheidungen. Eine allgemeine Kennzeichnungspflicht für KI-generierte Inhalte ist dagegen kaum zielführend.

Problemstellung

Die Gewährleistung von Transparenz ist eine der **zentralen Herausforderungen** bei der Verwendung von Künstlicher Intelligenz (KI). Menschen sind sich seit jeher gewohnt, mit Menschen zu interagieren. Zudem bestand bisher die auf Erfahrung gestützte Erwartung, dass Entscheidungen grundsätzlich von Menschen getroffen und Texte, Bilder, Videos und andere Inhalte von Menschen geschaffen werden. Diese Erfahrungen und Erwartungen werden mit der Verwendung von KI zunehmend in Frage gestellt. Menschen interagieren nicht mehr nur mit Menschen, sondern auch mit Chatbots, Entscheidungen werden zunehmend von KI-Systemen getroffen und Texte, Bilder und Videos werden von generativen KI-Systemen geschaffen. Wie die Interaktionen erfolgen, die Entscheidungen getroffen und die Texte, Bilder und Videos erstellt werden, ist meist unklar. Damit stellt sich die Frage, ob, in welchen Fällen und mit welchen Mitteln bei der Verwendung von KI Transparenz geschaffen werden muss.

Transparenz ist kein Selbstzweck. Vielmehr dient sie in aller Regel einem bestimmten Zweck, bspw. dem Schutz vor Diskriminierung oder Manipulation oder der Durchsetzung von Rechtsansprüchen. Zudem soll Transparenz über die Verwendung von KI-Systemen das **Vertrauen** in diese Systeme und deren **Akzeptanz** erhöhen und in der Bevölkerung das Bewusstsein und Verständnis für diese Technologie fördern. Bisweilen wird Transparenz auch aus Gründen der Fairness gefordert oder als Mittel zur Stärkung der Autonomie, der Effizienz oder der Rechenschaft (*accountability*) verstanden. Um diese Zwecke zu erfüllen, ist das Schaffen einer umfassenden Transparenz über alle Aspekte eines KI-Systems (bspw. die Architektur, die verwendeten Trainingsdaten, etc.), in aller Regel weder nötig noch sinnvoll. Die meisten Menschen könnten mit einer solchen umfassenden Transparenz ohnehin nichts anfangen. Der Forderung nach Transparenz gegenüber den Anbietern (*provider*) und Betreibern (*deployer*) von KI-Systemen stehen zudem legitime Interessen am Schutz von Geschäftsgeheimnissen gegenüber. Die **Transparenz** ist deshalb in der Regel **beschränkt**.

Bei der Verwendung von KI-Systemen lassen sich verschiedene **Dimensionen von Transparenz** unterscheiden. Zum einen stellt sich die Frage, ob für Menschen bei der Interaktion mit einem KI-System (bspw. einem Chatbot) erkennbar sein muss, dass sie nicht mit einem Menschen, sondern mit KI interagieren; vergleichbare Fragen stellen sich bei Inhalten, die von einem generativen KI-System produziert worden sind («**Erkennbarkeit**»). Zum anderen stellt sich die Frage, ob und inwiefern Menschen verstehen müssen, wie das KI-System zu einer bestimmten Vorhersage, Empfehlung oder Entscheidung gelangt oder einen bestimmten Inhalt erstellt («**Erklärbarkeit und Nachvollziehbarkeit**»). Die

Transparenz bezieht sich hier auf die Funktionsweise eines KI-Systems und auf die verwendeten Trainings- und Inputdaten. Auch in zeitlicher Hinsicht lassen sich zwei Dimensionen unterschieden. Transparenz kann vor oder bei der Interaktion mit einem KI-System sichergestellt werden, bspw. durch eine vorgängige Information über den Einsatz eines Chatbots, sie kann aber auch im Nachhinein und nur auf Nachfrage gewährt werden, bspw. durch Offenlegungspflichten sowie Einsichts- und Auskunftsrechte.

Der Forderung nach Transparenz von KI-Systemen sind auch aus **technischen Gründen Grenzen** gesetzt. Es ist eine wesentliche Eigenschaft dieser Systeme, dass sie nicht von Menschen vorgegebenen Regeln folgen, sondern über das Erkennen von statistischen Zusammenhängen in Daten selbst Regeln entwickeln, um von einem bestimmten Input zu einem bestimmten Output zu gelangen. Seit einigen Jahren wird allerdings intensiv an **technischen Lösungen** gearbeitet, welche die Erklärbarkeit und Nachvollziehbarkeit von KI-Systemen erhöhen sollen. Unter dem Sammelbegriff «**Explainable AI**» hat sich eine neue Forschungsrichtung etabliert, die sich damit befasst, KI-Systeme besser erklär- und nachvollziehbar zu machen, bspw. sollen bei durch KI-Systeme getroffenen Entscheidungen die für die Resultate kausalen Faktoren ermittelt werden.

Herausforderungen und Lösungsansätze

Interaktion mit KI-Systemen

Menschen sind sich gewohnt, mit anderen Menschen zu interagieren. In vielen Konstellationen wird die **Interaktion mit einem Menschen** aber heute durch die **Interaktion mit einer Maschine** ersetzt. Viele Unternehmen verwenden Chatbots, um zumindest auf einer ersten Stufe mit ihren Kund:innen zu kommunizieren. Auch wenn diese Interaktionen (noch) nicht immer überzeugen, ist allein aufgrund der Kostenersparnis davon auszugehen, dass Chatbots rasch in immer mehr Konstellationen verwendet werden. Absehbar ist auch der zunehmende Einsatz von Bedienungsrobotern in Restaurants und Pflegerobotern in Spitälern und Heimen. Früher oder später wird zudem mit autonomen Fahrzeugen und Drohnen zu rechnen sein.

In vielen Konstellationen ist für Menschen ohne weiteres erkennbar, dass sie mit einer Maschine interagieren, so etwa bei Bedienungs- und Pflegerobotern. Das ist aber nicht immer gewährleistet, bspw. wenn ein Chatbot einen Namen trägt und mit dem (KI-generierten) Foto eines Menschen erscheint. Hier stellt sich die Frage, ob und gegebenenfalls wie sicherzustellen ist, dass **Menschen erkennen**, dass sie nicht mit einem anderen Menschen, sondern **mit einem KI-System interagieren.** Für das Schaffen von Transparenz spricht, dass sie Vertrauen schaffen und die Akzeptanz von KI-Systemen erhöhen kann. Möglicherweise werden Menschen nicht mehr bereit sein, mit KI-Systemen zu interagieren, wenn sie bei einer früheren Interaktion mit einem solchen System meinten, es mit einem Menschen zu tun zu haben. Für viele Menschen mag es in manchen Konstellationen irrelevant sein, ob sie mit einem Menschen oder einer Maschine interagieren. Es ist aber denkbar, dass Menschen Maschinen ganz grundsätzlich nicht vertrauen und nicht mit diesen interagieren wollen und sie deshalb ein ausgeprägtes Bedürfnis nach Transparenz haben. Gegen das Schaffen von Transparenz bei der Interaktion mit einem KI-System spricht wenig. Relevant ist einzig der (allerdings in aller Regel geringe) Aufwand, den die Verwender von KI-Systemen leisten müssen, um die Erkennbarkeit zu gewährleisten. In vielen Fällen wird sich die Interaktion im Sinn von «**transparency by design**» mit wenig Aufwand so gestalten lassen, dass für die betroffenen Menschen ohne weiteres erkennbar ist, dass sie mit einem KI-System interagieren, bspw. indem sich ein Chatbot vor dem Start einer Konversation als solcher zu erkennen gibt.

Entscheidungen von KI-Systemen

Entscheidungen wurden bisher grundsätzlich von Menschen getroffen. Das gilt jedenfalls für komplexe Entscheidungen und solche, die für die betroffenen Personen eine gewisse Relevanz haben. KI-Systeme sind heute aber in der Lage, auch komplexe Entscheidungen zu treffen, die für die Betroffenen eine grosse Relevanz haben, bspw. über die Gewährung eines Kredits oder die Einstellung oder Entlassung einer Mitarbeitenden. Dass KI-Systeme in der Lage sind, komplexe Entscheidungen zu fällen, sagt allerdings noch nichts über die Qualität dieser Entscheidungen aus. In gewissen Konstellationen erreicht oder übertrifft die Qualität der Entscheidungen von KI-Systemen diejenige von Menschen (bspw. bei der Vorhersage von Aktienkursen), in anderen sind Menschen nach wie vor überlegen (bspw. im Erkennen von kulturellen Nuancen und Tonalitäten bei Übersetzungen). In vielen Konstellationen lässt sich die Qualität durch ein Zusammenwirken von Mensch und Maschine maximieren (bspw. in der medizinischen Diagnostik, insb. in der Radiologie). Bedenken bei der Qualität und eine grundsätzliche Skepsis gegenüber

automatisierten Entscheidungen können die **Akzeptanz der Entscheidungen von KI-Systemen** auf einer individuellen und gesellschaftlichen Ebene in Frage stellen

Wie bei der Interaktion mit KI-Systemen stellt sich auch bei Entscheidungen von KI-Systemen die Frage, ob für die betroffenen Personen erkennbar sein muss, dass eine Entscheidung von einem KI-System getroffen wurde (Erkennbarkeit). Bejaht man dies, stellt sich die Frage, ob nicht nur über den Umstand einer automatisierten Entscheidung, sondern auch darüber informiert werden muss, **wie die Entscheidung eines KI-Systems zustande gekommen ist** (Erklärbarkeit und Nachvollziehbarkeit). Die Transparenz über diese beiden Aspekte kann helfen, die Qualität der Entscheidungen zu prüfen und die Akzeptanz von automatisierten Entscheidungen zu erhöhen. Sind die Bedenken hinsichtlich der Qualität der Entscheidungen zu gross oder die Entscheidungen zu wichtig, ist es auch denkbar, Entscheidungen von KI-Systemen in gewissen Bereichen oder gar allgemein zu verbieten.

KI-generierte Inhalte

Generative KI-Systeme sind in der Lage, innert Bruchteilen von Sekunden verschiedenartige Inhalte zu produzieren, von Text und Bild über Musik bis hin zu Videos. Die Qualität dieser Inhalte hat sich in jüngerer Zeit stark verbessert und die Kosten für deren Erstellung sind marginal. Generative KI-Systeme werden schon heute in beruflichen und privaten Konstellationen für zahlreiche Zwecke verwendet. Es ist davon auszugehen, dass mit einer weiteren Verbesserung der Systeme auch deren Nutzung weiter zunehmen wird. Besonders problematisch ist, wenn generative KI-Systeme für das Erstellen von sog. *deep fakes* verwendet werden.

Schon heute ist oft nicht mehr erkennbar, ob ein Inhalt von einem Menschen oder einem KI-System geschaffen wurde. Damit stellt sich die Frage, ob und gegebenenfalls wie die Verwender von KI-Systemen erkennbar machen müssen, **ob ein Inhalt von einem Menschen oder einem KI-System erstellt wurde**. Ob eine allgemeine Pflicht zur Kennzeichnung von KI-generierten Inhalten geschaffen werden soll, ist allerdings fraglich, zumal es für viele Menschen in vielen Fällen nicht darauf ankommen wird, wie ein Inhalt erstellt wurde. Vielmehr ist zu erwarten, dass sich die Menschen mit der zunehmenden Verwendung von generativen KI-Systemen an automatisiert erstellte Inhalte gewöhnen und **nicht (mehr) davon ausgehen werden, dass Bilder oder Videos die Realität wiedergeben**. Eine allgemeine Kennzeichnungspflicht dürfte deshalb, wenn überhaupt, nur in einer Übergangsphase sinnvoll sein. Denkbar ist aber, dass in **gewissen Bereichen** auch längerfristig erkennbar sein muss, ob Inhalte von einem Menschen oder einem KI-System geschaffen wurden, bspw. bei Inhalten von Medien. Wenn man eine Kennzeichnungspflicht für KI-generierte Inhalte schaffen will, muss man zudem definieren, ab wann ein Inhalt als KI-generiert zu qualifizieren ist, bzw. umgekehrt, welches Mass an menschlichem Einfluss genügen soll, damit ein Inhalt nicht mehr als KI-generiert gilt.

Wesentlich wirkungsvoller und wichtiger als eine allgemeine Kennzeichnungspflicht für KI-generierte Inhalte ist eine umfassende **Aufklärung und Sensibilisierung der Bevölkerung** für die Möglichkeiten und Gefahren von generativer KI. Namentlich muss sichergestellt werden, dass Bürger:innen und Konsument:innen verstehen, dass nicht nur Texte, sondern auch Bilder und Videos keine zuverlässigen Abbilder der Realität (mehr) sind. Der Prozess der Aufklärung und Sensibilisierung ist bereits im Gang, zumal das Problem in den klassischen Medien und in Social Media regelmässig und prominent thematisiert wird. Darüber hinaus können bereits bestehende und laufend weiter zu entwickelnde **technische Mittel** verwendet werden, um die Authentizität von Inhalten zu überprüfen.

Durchsetzung von Rechtsansprüchen

Die Durchsetzung von Rechtsansprüchen mittels Klage setzt voraus, dass der Kläger die Tatsachen beweisen kann, die einen bestimmten Tatbestand erfüllen. Wer bspw. einen Schadenersatzanspruch geltend macht, muss beweisen, dass ein Schaden entstanden ist, ein widerrechtliches Verhalten vorliegt und der Schaden durch das widerrechtliche Verhalten verursacht wurde (Kausalzusammenhang); besteht kein Vertragsverhältnis, muss der Kläger zudem das Verschulden des Schädigers nachweisen. Auch bei anderen Rechtsverletzungen, bspw. bei einer Diskriminierung oder einer Manipulation, muss der **Kläger den Beweis erbringen**, dass die Voraussetzungen des jeweiligen Tatbestands erfüllt sind. Das fällt oft nicht leicht.

Hat ein KI-System einen Schaden verursacht oder zu einer Diskriminierung geführt, kann die **Beweisführung**

besonders schwierig sein, weil für den Kläger in aller Regel nicht erkennbar ist, warum das KI-System eine bestimmte Empfehlung gemacht oder eine bestimmte Entscheidung gefällt hat, warum also bspw. ein autonomes Fahrzeug beschleunigt und nicht gebremst oder ein Empfehlungssystem einen Bewerber als ungeeignet qualifiziert hat. Ähnliche Probleme bestehen bei einem Verstoss gegen die Vorgaben des Datenschutzrechts und bei der Verletzung von Urheberrechten. Dort liegt die grösste Herausforderung im Nachweis, dass ein KI-System mit bestimmten Personendaten oder urheberrechtlich geschützten Werken trainiert wurde und dass diese Daten bzw. Werke im trainierten Modell noch vorhanden sind. Bisweilen lassen sich die relevanten Tatsachen allerdings allein aufgrund des **Outputs eines KI-Systems** nachweisen, ohne dass Informationen über die Funktionsweise des Systems oder die bearbeiteten Daten erforderlich wären. In bestimmten Fällen sieht das geltende Recht zudem massgebliche Erleichterungen vor, so namentlich bei Kausalhaftungen (bspw. bei der Produktehaftung), bei denen das Verschulden nicht nachgewiesen werden muss.

Trotz dieser Einschränkungen besteht in vielen Konstellationen die **Gefahr, dass Kläger ihre Rechtsansprüche nicht durchsetzen können**, weil sie nicht nachzuweisen vermögen, dass ein Schaden, eine Diskriminierung oder eine Manipulation durch ein KI-System verursacht oder das System mit ihren Personendaten oder urheberrechtlich geschützten Werken trainiert wurde. Das ist problematisch, weil das geltende Recht keinen (ausreichenden) Schutz mehr bietet, wenn die Durchsetzung der Ansprüche an der Beweisführung scheitert. Zwar stehen gewisse **technische Mittel** zur Verfügung, mit denen sich bspw. prüfen lässt, ob ein KI-System mit bestimmten Daten oder Werken trainiert wurde oder eine Diskriminierung vorliegt. Darüber hinaus muss aber auch die **Rechtsordnung** sicherstellen, dass von einer Rechtsverletzung Betroffene über die Informationen verfügen, die sie für die Durchsetzung ihrer Rechtsansprüche benötigen.

Rechtliche Erfassung

Informationspflicht bei Interaktion mit KI-Systemen

Um Transparenz über die **Interaktion mit einem KI-System** zu gewährleisten, sollte eine allgemeine Informationspflicht geschaffen werden, welche die Betreiber von KI-Systemen zur **Information über die Verwendung eines KI-Systems verpflichtet**. Die KI-Konvention des Europarates und die KI-Verordnung der EU sehen eine solche Informationspflicht vor (Art 15 Ziff. 2 KI-Konvention; Art. 52 Ziff. 1 KI-VO). In vielen Fällen wird es allerdings nicht erforderlich sein, die betroffenen Personen aktiv (bspw. durch einen Hinweis in Textform, ähnlich wie bei «Cookie Banners») zu informieren, weil die Verwendung eines KI-Systems schon aus den Umständen ersichtlich ist. Ist das nicht der Fall, muss die Interaktion mit einem KI-System bewusst so gestaltet werden, dass die Erkennbarkeit gewährleistet ist («**transparency by design**»). Nur wenn auch dies nicht möglich oder aus Sicht des Betreibers nicht wünschenswert ist, muss dieser die Erkennbarkeit aktiv durch die Vermittlung einer entsprechenden Information sicherstellen. Um das Ziel dieser Informationspflicht zu erreichen, muss der Betreiber nur erkennbar machen, dass ein KI-System verwendet wird (Erkennbarkeit), er muss aber nicht über dessen Funktionsweise informieren (Erklärbarkeit und Nachvollziehbarkeit).

Bei der Interaktion mit einem KI-System wird dieses in der Regel Daten über die betroffene Person bearbeiten. Bei der Bearbeitung von Personendaten durch Private sind die Vorgaben des **Datenschutzgesetzes (DSG)** einzuhalten. Dieses sieht neben dem allgemeinen Grundsatz der Transparenz (Art. 6 Abs. 3 DSG) Informationspflichten vor (Art. 19 f. DSG) und gewährt den betroffenen Personen ein weitgehendes Auskunftsrecht (Art. 25 ff. DSG). Eine Informationspflicht bei der Interaktion mit einem KI-System kann an sich direkt auf diese Bestimmungen gestützt werden. Das entspricht auch der Auffassung des Eidgenössischen Datenschutz- und Öffentlichkeitsbeauftragten (EDÖB). Dennoch könnte es sinnvoll sein, eine **Informationspflicht bei der Interaktion mit einem KI-System ausdrücklich im DSG zu regeln**, bspw. im Kontext der bestehenden Informationspflicht bei automatisierten Einzelentscheidungen (Art. 21 DSG). Für eine ausdrückliche Regelung spricht nicht nur die Rechtssicherheit, sondern auch, dass es in gewissen Konstellationen möglich (und datenschutzrechtlich sinnvoll oder gar erforderlich) sein kann, die Interaktion mit einem KI-System so auszugestalten, dass keine Personendaten bearbeitet werden (bspw. bei einem Bedienungsroboter). In diesen Fällen liesse sich die Informationspflicht nicht mehr auf die bestehenden Normen des DSG stützen, wohl aber auf eine ausdrückliche Informationspflicht, die alle Interaktionen mit einem KI-System erfassen könnte. Dass dann eine im DSG normierte Pflicht auch Konstellationen erfassen würde, in denen keine Personendaten bearbeitet werden, ist zwar gesetzessystematisch

unbefriedigend. Da kein anderer Erlass für das Einfügen einer solchen allgemeingültigen Regelung ersichtlich ist, kann (und sollte) das aber hingenommen werden. Das gilt umso mehr, als die Informationspflicht bei der Interaktion mit einem KI-System der bereits im DSG geregelten Informationspflicht bei automatisierten Einzelentscheidungen (Art. 21 DSG) nahesteht.

Informationspflicht bei Entscheidungen von KI-Systemen

Das DSG sieht eine Informationspflicht bei automatisierten Einzelentscheidungen vor, die auch bei Entscheidungen von KI-Systemen gilt. Nach Art. 21 Abs. 1 DSG muss der Verantwortliche die betroffene Person über eine Entscheidung informieren, die ausschliesslich auf einer automatisierten Bearbeitung von Personendaten beruht, wenn die Entscheidung für die betroffene Person mit einer Rechtsfolge verbunden ist oder sie erheblich beeinträchtigt. Keine Informationspflicht besteht, wenn die Entscheidung von geringer Tragweite ist oder nicht vollautomatisiert erfolgt. Das ist der Fall, wenn KI-Systeme als sog. «**decision support systems**» verwendet werden.

Sind die Voraussetzungen hinsichtlich Art und Relevanz der Entscheidung erfüllt, gibt die Bestimmung den betroffenen Personen das Recht, auf Antrag ihren **eigenen Standpunkt darzulegen** und zu verlangen, dass die automatisierte Einzelentscheidung **von einem Menschen überprüft** wird (Art. 21 Abs. 2 DSG). Die Regelung geht damit über eine blosse Informationspflicht hinaus. Die Informationspflicht und die damit verbundenen Rechte bestehen allerdings nicht, wenn die automatisierte Entscheidung in unmittelbarem Zusammenhang mit dem Abschluss oder der Abwicklung eines Vertrags steht und dem Begehren der betroffenen Person stattgegeben wird (Art. 21 Abs. 3 lit. a DSG) oder die betroffene Person ausdrücklich eingewilligt hat, dass die Entscheidung automatisiert gefällt wird (Art. 21 Abs. 3 lit. b DSG).

Neben der Informationspflicht dient auch das **Auskunftsrecht** des DSG der Transparenz. Dieses gibt den betroffenen Personen das Recht, vom Betreiber eines KI-Systems Informationen über die **Logik** zu verlangen, auf der eine automatisierte Einzelentscheidung beruht (Art. 25 Abs. 2 lit. f DSG). Anzugeben sind dabei die Kriterien, nach denen eine automatisierte Entscheidung gefällt wurde und die Personendaten, auf denen die Entscheidung beruht. Über die bei der Entscheidung verwendeten Algorithmen muss grundsätzlich nicht informiert werden. Bei KI-Systemen wird man aber verlangen, dass den Betroffenen auch die dem System (allenfalls) vorgegebenen Ziele und die Datenkategorien mitgeteilt werden, mit denen das System trainiert wurde.

Der **Gehalt** dieser Regelung erscheint **angemessen und ausreichend**, ihr **Anwendungsbereich** ist aber **zu eng**. Denn die von Art. 21 DSG adressierten Bedürfnisse bestehen nicht nur, wenn eine Entscheidung «ausschliesslich auf einer automatisierten Bearbeitung beruht», also vollständig automatisiert gefällt wird, sondern auch bei weitgehend automatisierten Entscheidungen. Im Ergebnis besteht zwischen diesen beiden Konstellationen kaum ein Unterschied, weil der Einsatz von sog. «**decision support systems**» in der Regel dazu führt, dass Menschen die von der Maschine vorgeschlagene Entscheidung akzeptieren und nicht mehr hinterfragen, bspw. weil die Zeit für eine Überprüfung nicht reicht oder weil sich die menschlichen Entscheider rechtfertigen müssen, wenn sie vom Vorschlag der Maschine abweichen. Der Anwendungsbereich der Bestimmung sollte deshalb auf **weitgehend automatisiert getroffene Entscheidungen** erweitert werden. Nicht sinnvoll wäre dagegen, die Regelung des DSG durch ein grundsätzliches (wenn auch durch Ausnahmen relativiertes) Verbot automatisierter Einzelentscheidungen zu ersetzen, wie es in Art. 22 DSGVO vorgesehen ist.

Kennzeichnungspflicht für KI-generierte Inhalte

Die KI-Konvention des Europarates sieht eine Kennzeichnungspflicht für KI-generierte Inhalte vor, verlangt aber keine umfassende, sondern nur eine den jeweiligen Kontexten und Risiken angemessene Transparenz (Art. 8 KI-Konvention). Die KI-Verordnung der EU verlangt, dass die Nutzer von KI-Systemen bei *deep fakes* offenlegen müssen, dass die Bild-, Ton- oder Videoinhalte künstlich erzeugt oder manipuliert wurden (Art. 52 Ziff. 3 KI-VO). Eine allgemeine Kennzeichnungspflicht erscheint allerdings, wie vorne ausgeführt, aus einer Reihe von Gründen wenig sinnvoll. Zudem dürfte eine solche Pflicht gerade in den Fällen wenig helfen, in denen sie besonders relevant wäre, namentlich bei der Erzeugung von «*deep fakes*». Denn wer Menschen mit einem «*deep fake*» täuschen will, wird auch dann nicht über die Verwendung eines KI-Systems informieren, wenn die Rechtsordnung eine solche Pflicht vorsieht, und/oder Mittel finden, um die vom Anbieter eines KI-Systems implementierte Kennzeichnung zu entfernen oder unkenntlich zu machen. Hinzu kommen relevante Abgrenzungsfragen und grosse Herausforderungen bei der Umsetzung: Da Inhalte bei der Verwendung von generativer KI regelmässig im Zusammenwirken von Mensch und Maschine entstehen,

lässt sich kaum sagen, wann ein Inhalt als «KI-generiert» zu qualifizieren und damit als solcher zu kennzeichnen ist. Wie eine umfassende Kennzeichnungspflicht sinnvoll umgesetzt werden könnte, ist damit weitgehend unklar. Schon heute zeichnet sich ab, dass der Aufwand der Umsetzung den Nutzen einer allgemeinen Kennzeichnungspflicht überwiegen würde.

Anstelle einer allgemeinen Kennzeichnungspflicht für KI-generierte Inhalte könnten **sektorspezifische Kennzeichnungspflichten** sinnvoll sein. Namentlich wäre es denkbar, eine Kennzeichnungspflicht für **automatisiert erstellte journalistische Inhalte** einzuführen. Da die Glaubwürdigkeit der Inhalte für den Erfolg der (meisten) traditionellen Medien zentral ist, haben sich viele Medien bereits selbst Regeln zum Umgang mit (generativer) KI gegeben. Zudem hat der Schweizer Presserat einen Leitfaden für KI im Journalismus erlassen, der die relevanten Fragen regelt. Dieser sieht unter anderem vor, dass mithilfe von KI erstellte Inhalte als solche zu kennzeichnen sind. Da die Selbstregulierung der Medien durch den Presserat in der Schweiz gut funktioniert, besteht kein Anlass, diese Frage gesetzlich zu regeln. Sollte sich künftig zeigen, dass in anderen Sektoren ein Bedarf nach Kennzeichnung von KI-generierten Inhalten besteht, wäre zu prüfen, ob das Ziel auch dort durch Selbstregulierung erreicht werden kann, oder ob der Gesetzgeber eine sektorspezifische Kennzeichnungspflicht vorsehen sollte.

Durchsetzung von Rechtsansprüchen

Die Durchsetzung von Rechtsansprüchen bei der Verwendung von KI-Systemen ist oft mit Schwierigkeiten verbunden. Das ist aber nicht immer der Fall. Nicht selten lassen sich die relevanten Tatsachen allein aufgrund des Outputs eines KI-Systems nachweisen, ohne dass Informationen über das Funktionieren des Systems oder die bearbeiteten Daten erforderlich wären, bspw. bei Manipulation. Bisweilen sieht auch das materielle Recht massgebliche Erleichterungen vor, etwa bei Kausalhaftungen, bei denen das Verschulden nicht nachgewiesen werden muss.

Die Anforderungen an die Beweisführung werden in der Schweizerischen Zivilprozessordnung (ZPO) geregelt. Diese beruht auf dem **Grundsatz der freien Beweiswürdigung** durch die Gerichte (Art. 157 ZPO). Auch wenn im Grundsatz ein strikter Beweis zu erbringen ist, genügt es, wenn das Gericht am Vorliegen einer Tatsachenbehauptung **keine ernsthaften Zweifel** mehr hat oder allenfalls verbleibende Zweifel als leicht erscheinen; ist ein strikter Beweis nicht möglich, etwa beim Kausalzusammenhang, genügt bereits der Nachweis der überwiegenden Wahrscheinlichkeit. Dieser flexible Ansatz ermöglicht Gerichten, dem Umstand Rechnung zu tragen, dass Kläger die Funktionsweise von KI-Systemen in der Regel nur beschränkt nachvollziehen und erst recht nicht strikt beweisen können. Das materielle Recht sieht zudem in gewissen Konstellationen **Beweiserleichterungen** (bspw. im Gleichstellungsgesetz) oder gar eine **Beweislastumkehr** (bspw. beim Verschulden bei vertraglicher Haftung) vor. Diese Mittel können auch beim Nachweis von Rechtsverletzungen im Zusammenhang mit KI-Systemen sinnvoll sein, insb. bei Diskriminierungen. Ein allgemeines Gleichbehandlungsgesetz könnte bspw. vorsehen, dass die betroffenen Personen die Anknüpfung an ein geschütztes Merkmal nur glaubhaft machen und die beklagte Partei das Vorliegen einer qualifizierten Rechtfertigung beweisen müsste. Zudem könnten die Möglichkeiten der Rechtsdurchsetzung durch eine Stärkung des kollektiven Rechtsschutzes verbessert werden, insb. durch die geplante Revision des Verbandsklagerechts.

Das **Auskunftsrecht des DSG** ermöglicht betroffenen Personen, von den Verantwortlichen umfassende Informationen über die Bearbeitung der sie betreffenden Personendaten zu erhalten (Art. 25 DSG). Die Durchsetzung von datenschutzrechtlichen Ansprüchen wird deshalb kaum je an der fehlenden Transparenz scheitern. Das Auskunftsrecht steht zwar grundsätzlich nur zur Verfügung, um die im DSG vorgesehenen Rechte geltend zu machen. Da das DSG aber dem Schutz der Persönlichkeit und der Grundrechte von natürlichen Personen dient (Art. 1 DSG), wäre es denkbar, dass betroffene Personen das datenschutzrechtliche Auskunftsrecht auch verwenden können, um Ansprüche wegen Diskriminierung oder Manipulation geltend zu machen, die sich als Persönlichkeitsverletzungen qualifizieren lassen. Ausgeschlossen erscheint eine Nutzung des Auskunftsrechts allerdings, wenn es um die Durchsetzung von Schadenersatzansprüchen geht.

Die ZPO kennt prozessuale Mitwirkungspflichten, namentlich die **Zeugnis- und Editionspflicht** (Art. 160 ZPO). Aufgrund der Editionspflicht können Prozessparteien und nicht am Prozess beteiligte Dritte zur Herausgabe von Urkunden verpflichtet werden. Das gilt auch für Dokumentationen über KI-Systeme. Auch wenn die Editionspflicht unter Umständen eingeschränkt werden kann, dürfte sich dieses Mittel in vielen Konstellationen nutzen lassen, um die für die

Durchsetzung von Rechtsansprüchen bei der Verwendung von KI-Systemen erforderlichen Informationen zu erhalten. Weitere **Offenlegungspflichten**, wie sie die EU in der KI-Haftungsrichtlinie vorsieht (Art. 3 KI-Haftungs-RL), scheinen deshalb nicht erforderlich. Das gilt auch für die in der KI-Verordnung vorgesehenen, teilweise sehr weit gehenden **Dokumentationspflichten** (Art. 11 f. KI-Verordnung). Statt für eine grosse Vielzahl sehr unterschiedlicher Konstellationen gesetzliche Dokumentationspflichten zu schaffen, die in vielen Fällen unnötig und in anderen ungenügend sein werden, sollte es grundsätzlich den Anbietern und Betreibern von KI-Systemen überlassen werden, jeweils selbst zu prüfen, welche Dokumente sie erstellen müssen, um den datenschutzrechtlichen Anforderungen an die Transparenz genügen zu können und beim Eintritt einer Rechtsverletzung nachweisen zu können, dass ein Schaden, eine Diskriminierung oder eine andere Rechtsverletzung nicht durch einen Fehler des KI-Systems verursacht worden ist. Zumindest in bestimmten Konstellationen (bspw. bei Diskriminierung) könnte eine **Beweislastumkehr** starke Anreize für das Erstellen ausreichender Dokumentationen setzen. Darüber hinaus kann es sinnvoll sein, in **sektorspezifischen Regulierungen** Dokumentationspflichten vorzusehen, bspw. bei Medizinprodukten. In erster Linie sollten die erforderlichen Dokumentationen aber in **technischen Standards** geregelt werden. Diese haben gegenüber gesetzlichen Dokumentationspflichten den Vorzug, dass sie spezifischer gefasst und laufend der technischen Entwicklung angepasst werden können.

Impressum

Herausgeberin:
Center for Information Technology,
Society, and Law (ITSL)
Universität Zürich
Rämistrasse 74|38
8001 Zürich