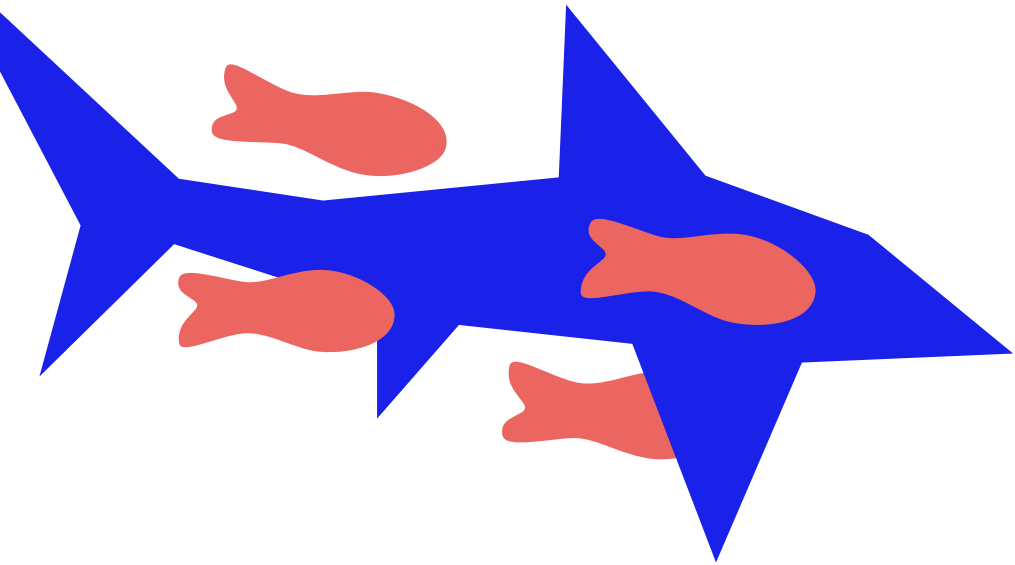
Der menschliche Umgang mit Künstlicher Intelligenz und deren Regulierung

Thomas Burri
Juliane Beck
François Fleuret
Serhiy Kandul
Markus Kneer
Vincent Micheli
Markus Christen



Professor Dr. iur. Thomas Burri, LLM., RA,
Professor für internationales Recht und Europäisches Recht, Universität St. Gallen

Juliane Beck,
Doktorandin in Recht, Universität St. Gallen

Dr. François Fleuret,
Professor für Informatik (maschinelles Lernen), Universität Genf

Dr. Serhiy Kandul,
Post-Doc in Verhaltensökonomie, Universität Zürich

Professor Dr. Markus Kneer,
Professor für Ethik der Künstlichen Intelligenz, Universität Graz

Vincent Micheli,
Doktorand in Informatik (maschinelles Lernen), Universität Genf

PD Dr. Markus Christen,
Geschäftsführer der Digital Society Initiative, Universität Zürich

Die Technologie der Künstlichen Intelligenz (KI) entwickelt sich zurzeit rasend schnell. Die Behörden reagieren auf diese Entwicklung mit entsprechenden Regulierungsvorhaben, doch zahlreiche Unsicherheiten verkomplizieren diese Vorhaben.

Die Zahl und Vielfalt an KI-Systemen nimmt immer weiter zu. Hochspezifische Anwendungen halten Einzug in zahlreiche, bereits separat regulierte Spezialgebiete. KI selbst ist zudem vielseitig einsetzbar. Bei vielen KI-Anwendungen stellt sich die Frage, welche Rolle die Person spielen soll, die mit der KI mehr oder weniger direkt in Kontakt steht.

Ein vom Schweizerischen Nationalfonds gefördertes Projekt hat in den vergangenen Jahren einen entscheidenden Aspekt dieser Frage untersucht, nämlich, wie und in welchem Ausmass der Mensch Kontrolle über KI-Anwendungen ausüben könnte und sollte. Damit verbunden wurden in einem separaten Projekt auch die zurzeit geeignetsten Formen der Regulierung von KI untersucht.

Die vorliegende Broschüre fasst die wissenschaftlichen Ergebnisse in der Form von Empfehlungen zusammen.

Diese richten sich an die Politik, die Behörden, sowie die Allgemeinheit.

Menschliche Kontrolle und menschliche Aufsicht

Die Idee, KI-Anwendungen menschlicher Kontrolle zu unterstellen, wurde bereits vor zehn Jahren prominent geäussert, als die Staaten auf internationaler Ebene (im Rahmen der «Convention on Certain Conventional Weapons Systems», CCW) den Versuch unternahmen, sogenannte autonome Waffensysteme (auch genannt «Killer Robots») zu regulieren. Das in Genf ansässige Gremium konnte sich jedoch nicht einigen, welches Mass an menschlicher Kontrolle über die in Rede stehenden, zunehmend auf KI gestützten Waffensysteme rechtlich erforderlich ist.

Indes hat die Forderung nach *menschlicher Kontrolle über KI* in aktuellen Regulierungsvorhaben Berücksichtigung gefunden. Der mittlerweile in Kraft getretene Rechtsakt der Europäischen Union zur Regulierung von KI, die sog. KI-Verordnung (auch: «KI Akt» oder «KI Gesetz», auf Englisch «EU AI Act») basiert grundlegend auf der Idee menschlicher Kontrolle über KI (findet allerdings auf Waffensysteme keine Anwendung).

Insbesondere beinhaltet er eine Vorschrift zu menschlicher *Aufsicht* über KI-Anwendungen, welche in die «Hochrisiko-Kategorie» fallen (siehe Art. 14 der KI Verordnung).

Sowohl bei der weiter gefassten Kontrolle über KI, als auch bei der stärker auf die konkrete Mensch-Maschine-Interaktion bezogene KI-Aufsicht dreht sich die Diskussion letztlich um das Zusammenwirken zwischen Mensch und Maschine.

Hintergrund dieses Regulierungsansatzes ist, dass der Mensch die KI typischerweise in der einen oder anderen Form steuert oder beherrscht und deswegen auch für das Ergebnis (Output) (rechtlich) verantwortlich zu sein scheint.Vorrichtungen, die das Erfordernis der Kontrolle bzw. Aufsicht über KI theoretisch umsetzen könnten, umfassen typischerweise Haltemechanismus («stop button») oder eine stichprobeweise Beobachtung des Verhaltens des KI-Systems.

Allerdings können Formen menschlicher Kontrolle bzw. Aufsicht über KI weiter ausdifferenziert sein. Zu denken ist beispielsweise an Erfordernisse wie Transparenz, Datenverwaltung, oder Risiko-Management, mithilfe derer die Gefahren von KI-Anwendungen eingehegt werden sollen. Die KI-Verordnung der EU schreibt dies bereits für besonders gefährliche (d.h., «Hochrisiko») KI-Systeme vor.

Die vorliegende Broschüre und die Forschung, auf die sie sich stützt, konzentrieren sich indes weitgehend auf konkrete Interventionsmöglichkeiten in Bezug auf laufende, KI-gestützte Prozesse.

Heisse und kalte Künstliche Intelligenz

Für ein grundlegendes Verständnis von Kontrolle über KI ist es hilfreich, zwischen «heisser» und «kalter» KI zu unterscheiden. «Heiss» ist die KI, wenn es sich um einen laufenden Interaktionsprozess handelt, bei dem Zeit- und Ressourcendruck besteht und der Mensch teils sekundenschnell eine Entscheidung über eine mögliche Intervention in einen KI-gestützten Prozess treffen muss.

Dies ist beispielsweise der Fall, wenn ein Fahrzeug gesteuert wird oder ein Raketenangriff abgewehrt werden soll.

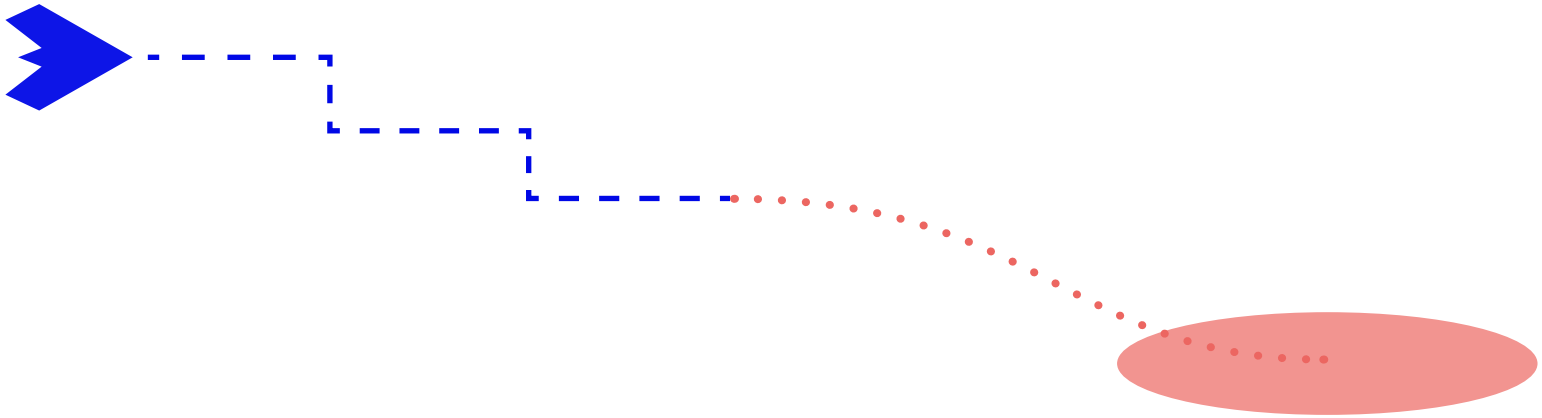
Bei «kalter» KI stehen demgegenüber ausreichend Zeit und Ressourcen zur Verfügung, um den Output der KI zu überprüfen. So kann zum Beispiel die Polizei ausreichend Zeit und Personal haben, um einen KI-generierten Output zu hinterfragen und verifizieren (bspw. die Identifikation einer verdächtigen Person auf Grundlage der Analyse und Auswertung biometrischer Daten). Auch kann z.B. der Output eines sog. «Large Language Models» mittels Recherche überprüft werden. Es sei jedoch darauf hingewiesen, dass die beiden Kategorien «heiss» und «kalt» primär didaktischen Zwecken dienen, denn in der KI-Praxis verschwimmen sie häufig. Die in der vorliegenden Broschüre verbriefte Forschung beschäftigt sich primär mit der Kontrolle über «heisse» KI.

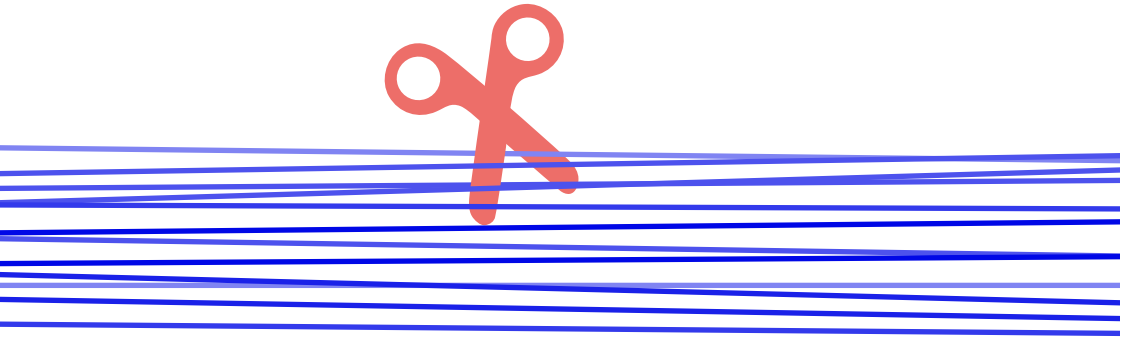
Landung mit Hilfe von Künstlicher Intelligenz

Die Forschungsergebnisse wurden mit Hilfe von sozialwissenschaftlichen Experimenten erzielt. Zahlreiche Testpersonen hatten dabei am Computer im Labor die Aufgabe, ein fliegendes Objekt mit der Hilfe von KI sicher zu landen, wobei das Ausmass der Unterstützung je nach Fähigkeitslevel der Testpersonen abgestuft war.

Der Prozess war «heiss», denn die Testpersonen hatten für ihre Landungen nur eine beschränkte Zeit zur Verfügung und waren auf sich allein gestellt.

Um einen realitätsgetreuen Prozess zu schaffen, wurde eine für die entsprechende Umgebung gängige KI besonders trainiert und fein abgestimmt. Dadurch wurde eine Situation geschaffen, welche reale Prozesse (wie etwa die Landung einer Raumkapsel auf dem Mond oder einer Drohne auf einem Flugzeugträger, bzw. die Steuerung eines Autos) abbildet. Die Gültigkeit, bzw. Validität der Forschungsergebnisse ist beschränkt auf solche Sachverhalte.





Empfehlungen:

Kontroll- oder Aufsichtspflichten über Künstliche Intelligenz sollten anwendungsbezogen konkretisiert werden

Es empfiehlt sich, das Erfordernis der menschlichen Kontrolle bzw. Aufsicht über KI in Bezug auf konkrete KI-Anwendungen zu formulieren. Wird bloss eine allgemein gehaltene Pflicht der Betreiberin normiert, menschliche Aufsicht über KI zu garantieren, droht diese Pflicht ins Leere zu laufen bzw. nicht den gewünschten Effekt zu erzielen.

Grund dafür ist zum einen, dass KI-Systeme oft nicht nur hochkomplex, sondern auch sehr unterschiedlich in ihrer Konfiguration und Funktionsweise sind, sodass generische Anforderungen ohne Berücksichtigung der technischen Feinheiten und des Anwendungskontextes wenig zielführend sind. Zum anderen behielte die Betreiberin im Falle unspezifisch gehaltener Regulierung ein ungebührliches Mass an Freiheit bei der Umsetzung der Kontroll- bzw. Aufsichtspflicht. Wirksame menschliche Kontrolle bzw. Aufsicht über KI wäre folglich nicht gewährleistet, sodass die Letztverantwortlichkeit und auch die rechtliche Haftung für einen möglicherweise fehlerhaften KI-Output schwerlich bei dem mit dem KI-System interagierenden Entscheidungsträger verortet werden könnte. Eine generisch formulierte Kontroll- bzw. Aufsichtspflicht wäre damit praktisch und rechtlich bedeutungslos.

Daher ist zu fordern, dass die Pflicht spezialgesetzlich, d.h. in Bezug auf Anwendungen in einem konkreten Umfeld – z.B. in der Medizin, im Strassenverkehr, im Finanzbereich oder der Justiz – ausgestaltet wird.

Quellen (siehe Literaturverzeichnis am Ende):
(1), (2), (4), (5), (6).

Menschliche Eingriffsmöglichkeiten in laufende, KI-gestützte Prozesse sind kein Allheilmittel

Intuitiv mögen menschliche Eingriffsmöglichkeiten in kontinuierliche, d.h. laufende bzw. «heisse», KI-gestützte Prozesse, sinnvoll erscheinen.

In der praktischen Anwendung erweisen sich solche Massnahmen allerdings häufig als wenig zielführend. Menschen können die Erfolgsaussichten und die Verlässlichkeit von KI-Systemen meist nur schlecht beurteilen. Oft über- oder unterschätzen Sie ihre eigenen Fähigkeiten oder vertrauen dem KI-generierten Output zu sehr («automation bias») bzw. auf ungerechtfertigte Weise zu wenig («algorithm aversion»). Die Aufmerksamkeitsspanne von Menschen ist in der Regel kurz. Da menschliche Entscheidungsträger jedoch in den allermeisten KI-gestützten Prozessen als letztes Glied in der Entscheidungskette fungieren (und sei es, dass sie eine schlichte Aufsichtsfunktion erfüllen), wird ihnen die Verantwortungsübernahme und Rechenschaftspflicht im Falle von fehler- oder schadhaftem Output aufgebürdet. Mit anderen Worten werden menschliche Entscheidungsträger somit im Falle von Fehlfunktionen des KI-Systems zu Sündenböcken degradiert.

Dabei ist auch zu berücksichtigen, dass Ausbildung und Training im Umgang mit spezifischen KI-Systemen die Erfolgschancen menschlicher Kontrolle (insb. Intervention) nicht zwangsläufig und in ausreichendem Masse erhöhen. Vor diesem Hintergrund sollten innerhalb des Haftpflichtregimes neue Lösungsansätze entwickelt werden.

Quellen: (2), (4), (6).

Die Erklärbarkeit hochleistungsfähiger KI-Systeme kann nicht ohne Weiteres angenommen werden

Es scheint, als bestünde eine Möglichkeit zur besseren Kontrolle von KI-Systemen darin, deren Erklärbarkeit zu erhöhen und dementsprechend eine Erklärungspflicht rechtlich zu normieren. Obwohl derzeit intensiv an der Erklärbarkeit von KI-Systemen geforscht wird («explainable AI», kurz: XAI), bleiben viele Systeme mit hoher Leistungsfähigkeit – insbesondere sog. «deep learning» Systeme, die mit hochkomplexen neuronalen Netzen arbeiten – weitgehend unerklärbar. Wenngleich der Input bekannt ist, so bleiben beispielsweise die Mechanismen, die einen konkreten Output generieren, aufgrund ihrer Komplexität unerklärbar. Das konkrete KI-System erscheint somit als eine sog. «Black Box».

Eine Erklärungs- bzw. Transparenzpflicht für KI kann sich deshalb nur auf die Input-Daten und/oder die grundsätzliche Modellgestaltung beziehen. Sie muss zudem an den Anwendungskontext und den Adressatenkreis angepasst sein. Mit Blick auf bestehende Mechanismen, welche darauf abzielen, die Vorhersehbarkeit von KI-Output zu erhöhen (sog. «confidence scores» oder «heat maps»), ist anzuraten, dass deren Wirksamkeit im konkreten Anwendungsfeld genau geprüft wird.

Letztlich gilt es zu bedenken, dass eine tiefgreifende, gesetzlich statuierte Erklärungspflicht den leistungsfähigsten KI-Systemen, insb. den «Large Language Models», einen Riegel vorschieben würde.

Quellen: (4), (7).

Eine umfassende Regulierung von Künstlicher Intelligenz wirkt sich deutlich auf Innovationsbereitschaft und -wille aus

Der schweizerische Gesetzgeber könnte sich anschicken, KI umfassend, d.h. in sämtlichen Anwendungsfeldern und -bereichen («horizontal»), zu regulieren.

Die EU hat sich für diesen Weg entschieden. Eine horizontale Regulierung von KI bringt jedoch auf absehbare Zeit ein hohes Mass an Unsicherheit für Unternehmen, insb. kleine und mittelständische, mit sich. Dies liegt vor allem daran, dass KI noch nie zuvor geregelt wurde und sich erst in Zukunft zeigen wird, ob sich insb. die EU KI Verordnung bewährt.

Die derzeitigen Unsicherheiten wirken sich indes auf absehbare Zeit negativ auf die Innovationsbereitschaft und den Innovationswillen von Unternehmen aus. Einen solchen Negativtrend kann eine später eintretende Rechtssicherheit kaum vollständig kompensieren.

Eine behutsame Anpassung der bestehenden Gesetzeslage in einzelnen Anwendungsbereichen brächte hingegen weniger Rechtsunsicherheit mit sich. Ferner ist zu bedenken, dass in der Schweiz und in Europa ansässige Unternehmen, die KI wirtschaftlich entwickeln und nutzen wollen (insbesondere Start-Ups), dies in aller Regel mit guten Absichten tun.

Gesetzgebung und Verwaltung sollten daher vorwiegend potenziell unseriöse Anwendungen von Künstlicher Intelligenz ins Auge fassen (etwa solche, bei denen unklar bleibt, ob sie überhaupt technisch funktionieren oder auf legal erlangten Daten beruhen).

Quellen: (1), (3).

Das bestehende Recht erfasst Künstliche Intelligenz bereits zu weiten Teilen

KI stellt uns in vielerlei Hinsicht vor neue Herausforderungen. Dennoch werden KI-Systeme durch das bestehende Recht bereits zu weiten Teilen erfasst. Insbesondere die Grundrechte, das Recht des Persönlichkeitsschutzes, das Datenschutzrecht, das Urheberrecht, das Haftungsrecht, das Arbeitsrecht und das Strafrecht finden auf KI-Systeme Anwendung und hegen diese zumindest teilweise ein.

Auf der Grundlage des bestehenden Rechts werden zudem Verwaltungsbehörden und Gerichte, aber auch Private viele der Probleme und Unsicherheiten, die sich bei der Anwendung des bestehenden Rechts auf KI ergeben, lösen können. KI ist also auch jetzt nicht völlig dereguliert und bewegt sich damit nicht in einem rechtlichen Vakuum.

Nichtsdestotrotz bringt KI zahlreiche spezifische Herausforderungen mit sich, welche die Gesetzgebung anpacken sollte. Allerdings muss sie dies nicht notwendigerweise in einem allumfassenden Gesetzgebungsakt tun.

Quellen: (2), (3), (5).

Künstliche Intelligenz hat soziale Auswirkungen, die separat von den Betriebsrisiken der KI angegangen werden sollten

KI wirkt sich auf die Arbeitswelt, das Bildungswesen, die soziale Sicherheit, auf Verteilungsfragen sowie das liberale Wirtschaftsmodell aus.

Vor diesem Hintergrund sehen sich die politischen Organe mit Entscheidungen unterschiedlicher Art konfrontiert.

1. Zum einen müssen sie festlegen, ob KI in gewissen Anwendungsfeldern oder -formen möglicherweise gänzlich unerwünscht ist und daher bereichsspezifisch verboten werden sollte.
2. Zum anderen müssen sie entscheiden, wie die mit KI einhergehenden Veränderungen, insbesondere jene sozialer Natur, begleitet werden sollen.
3. Ferner müssen sie die konkreten Betriebsrisiken von KI (wie z.B. versteckte Diskriminierungen oder die Beeinflussbarkeit von KI; also z.B. neuartige Formen von Cyberangriffen, die etwa auf der Ebene der Trainingsdaten einwirken) in den Blick zu nehmen und dafür konkrete regulatorische Lösungsansätze entwickeln.

Quellen: (1), (3), (5).

Der vielversprechendste Regulierungsansatz: Allgemeine Grundprinzipien mit spezialgesetzlichen Normen kombinieren

Die Anpassungen des positiven Rechts, die notwendig sind, um KI vernünftig zu erfassen und zu regulieren, sollten nicht im Wege eines umfassenden Gesetzgebungsaktes, d.h. eines „KI-Gesetzes“, vorgenommen werden.

Wie bereits angeklungen, wiese ein solcher einen zu hohen Abstraktionsgrad und damit einhergehend eine zu geringe Granularität auf.

Ferner zöge sich der Gesetzgebungsprozess übermässig in die Länge. Daher sollten zunächst lediglich abstrakte Grundsätze bzw. Leitlinien beschlossen werden, welche die allgemeinen Erwartungen an den Umgang mit KI-Systemen festlegen und somit insbesondere der Wirtschaft eine Richtschnur an die Hand geben.

Zu diesen Erwartungen zählen insbesondere der Schutz vor Diskriminierung (Bias), die Robustheit von KI-Systemen und ein klares Verantwortungsregime im Schadensfall. Die Bundesratsleitlinien von 2020 bezüglich KI weisen in die richtige Richtung und können für die Wirtschaft angepasst werden.

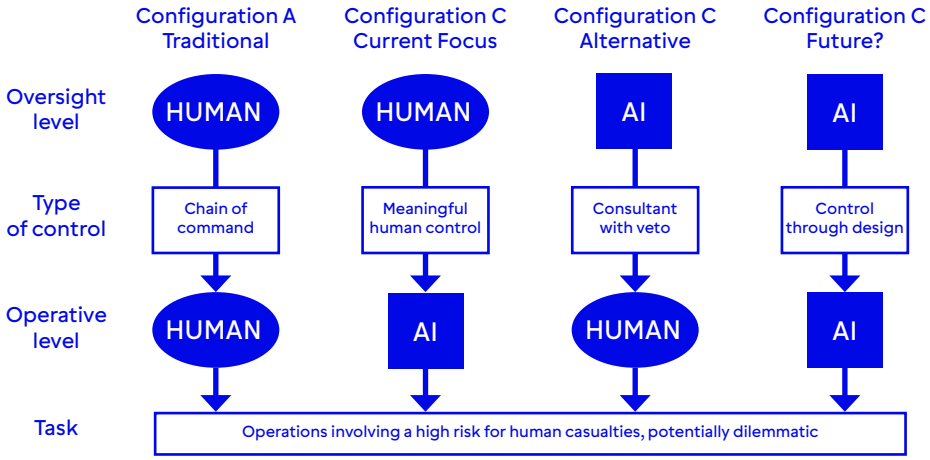
Im nächsten Schritt sollte die Gesetzgebung sodann damit beginnen, die Spezialgesetze gestaffelt nach Dringlichkeit anzupassen.

Quellen: (1), (6).

Die EU-Verordnung zu Künstlicher Intelligenz sollte in der Schweiz vorerst nicht explizit übernommen werden

Die Schweiz sollte zunächst davon absehen, die neue EU-Verordnung zur Regulierung von Künstlicher Intelligenz explizit zu übernehmen. Stattdessen sollten es die schweizerische Politik und Gesetzgebung jedem einzelnen Unternehmen überlassen, wann und inwieweit es sich nach der EU-Verordnung richten möchte. Dadurch wird in der Schweiz der Raum für Innovation bewahrt. Die Gesetzgebung sollte allerdings bei der Anpassung der schweizerischen Spezialgesetze Reibung mit der EU-Verordnung unbedingt vermeiden. Die Behörden sollten Aufklärung bezüglich der sich aus der EU-Verordnung ergebenden Pflichten leisten und den Unternehmen (insbesondere den Start-Ups) Unterstützung anbieten, um ihnen frühzeitig Compliance mit der EU-Verordnung und damit Zugang zum EU-Markt zu ermöglichen.

Quellen: (1).



Quelle: Markus Christen, Thomas Burri, Serhiy Kandul, Pascal Vöros, Who is controlling whom? Reframing “meaning human control” of AI systems in security, Ethics and Information Technology (2023) 25:10.

Literatur:

1. **Thomas Burri**, “A challenge for the law and artificial intelligence”, *Nature Machine Intelligence*, 5, 1508–1509 (2023), https://doi.org/10.1038/s42256-023-00768-5 (Paywall), Preprint: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4620764

2. **Markus Christen, Thomas Burri, Serhiy Kandul**, Pascal Vörös, “Who is controlling whom? Reframing ‘meaningful human control’ of AI systems in security”, (2023) *25 Ethics and Information Technology* (10)

3. Emmie Hine, Yasaman Yousefi, Parisa Osivand, Dirk Brand, Kholofelo Kugler, Pier Giorgio Chiara, “The AI Act Grand Challenge shows how autonomous robots will be regulated”, 8, *Science Robotics* 84 (2023), DOI: 10.1126/scirobotics.adk5632 (Paywall)

4. **Serhiy Kandul, Micheli Vincent, Juliane Beck, Thomas Burri, François Fleuret, Markus Kneer, Markus Christen**, “Human control redressed: Comparing AI and human predictability in a real-effort task”, (2023) 10 *Computers in Human Behaviour Reports* (May 2023, 100290), https://doi.org/10.1016/j.chbr.2023.100290

5. **Thomas Burri, Markus Christen, Juliane Beck,** Daniel Trusilo, “Eight Recommendations for Ethical and Legal Assessments of Robotic Systems Interacting with Humans”, in Woodrow Barfield, Ugo Pagallo, und Yueh-Hsuan Weng (Hrsg.), Research Handbook on the Law, Regulation, and Policy of Human Robot Interaction, CUP (im Erscheinen), Preprint: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4437806

6. **Juliane Beck, Thomas Burri**, “From ‘Human Control’ in International Law to ‘Human Oversight’ in the New EU Act on Artificial Intelligence, in Daniele Amoroso und Filippo Santoni De Sio (Hrsg.), Research Handbook on Meaningful Human Control of Artificial Intelligence Systems, Elgar, 2023 (im Erscheinen), Preprint: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4236554

7. **Serhiy Kandul, Vincent Micheli, Juliane Beck, Markus Kneer, Thomas Burri, François Fleuret, Markus Christen**, “Explainable AI: A Review of the Empirical Literature” (in Vorbereitung), Entwurf: https://ssrn.com/abstract=4325219

SNF Beitrag Nr. 407740_187494 / 1
Armasuisse W+T Vertrag Nr. 8003535283, 8003538711, 8003539311

Impressum:
Satz und Illustration: Malte Euler
Schrift: TOS Sans von ShowMeFonts
Papier: IGEPA Circle offset Premium White Recycle