Die Seite des SF | La page du FS

SF·FS

Tech Law Workshop

Angriffe auf KI-Systeme

Ein interdisziplinär zusammengesetztes Publikum aus Wissenschaft und Praxis traf sich am 10. November 2023 im Rahmen eines Tech Law Workshops, um diverse Aspekte von Systemen der Künstlichen Intelligenz (KI) und den Gefahren möglicher Angriffe auf diese Systeme zu diskutieren. Der in Zürich durchgeführte Workshop wurde veranstaltet vom Schweizer Forum für Kommunikationsrecht (SF-FS), dem Zentrum für Life Sciences-Recht (ZLSR) der Universität Basel und dem Center for Information Technology, Society, and Law (ITSL) der Universität Zürich.

Un public tout aussi diversifié et interdisciplinaire, issu de la science et de la pratique, s'est réuni le 10 novembre 2023 dans le cadre d'un Tech Law Workshop pour discuter de divers aspects des systèmes d'intelligence artificielle et des dangers d'éventuelles attaques contre ces systèmes. L'atelier était organisé à Zurich par le Forum suisse pour le droit de la communication (SF-FS), le Centre pour le droit des sciences de la vie (ZLSR) de l'Université de Bâle et le Center for Information Technology, Society, and Law (ITSL) de l'Université de Zurich.

I. Einführung

Der Tagungsleiter Prof. Dr. Alfred Früh, Professor für Privatrecht an der Universität Basel, begrüsste die Teilnehmerinnen und Teilnehmer und erläuterte das Thema und den Aufbau des Workshops. Die Tatsache, dass KI-Systeme auf vielfältige Weise angegriffen und manipuliert werden können, werde in der Industrie und der Rechtswissenschaft noch unterschätzt, zumal diese Angriffe im Gegensatz zu herkömmlichen Angriffen (wie *Hacking* oder *Phishing*) neue Merkmale aufwiesen. Inspiriert von einem Forschungsprojekt des ZLSR[1] sei es das Ziel des Workshops, das Problem und mögliche Lösungen interdisziplinär besser zu erfassen.

Im ersten Teil gehe es um die grundsätzliche Problematik von Angriffen auf KI-Systeme und darum, deren Vielfalt, Gefahren und Handlungsbedarf aufzuzeigen. Der zweite Teil konzentriere sich vor allem auf technische Massnahmen und das rechtliche Vorgehen gegen Angriffe auf KI. Dafür kämen in erster Linie das Strafrecht, das Haftungsrecht und das Immaterialgüterrecht in Frage.

Inesa Fausch, Dr. iur., Rechtsanwältin, Universität Basel.

Daniel Zeyer, MLaw, Universität Basel.

II. Erster Teil: Technische Grundlagen

1. Arten und Vorkommen von Angriffen auf KI-Systeme

Die Vortragsreihe eröffnete Dr. Kathrin Grosse, Wissenschaftlerin an der EPFL im Bereich der *Visual Intelligence for Transportation*. Mit ihrer Expertise in Computerwissenschaften und KI-Sicherheit erläuterte sie den Anwesenden, was ein Angriff auf ein KI-System ist, und nannte sowohl die verschiedenen Arten der Angriffe als auch die Folgen solcher Angriffe auf das KI-System. Sie unterschied dabei, ob sich der Angriff auf das KI-System und dessen Entwicklung an sich oder dessen potenziell gefährdete Komponenten, wie zum Beispiel die Trainings- oder Testdaten, konzentriert. Wie ein allfälliger Angriff aussehen kann, erklärte *Grosse* an einem konkreten Beispiel. Soll ein *Machine Learning Model (ML-Model)* Patienten mit gewissen Merkmalen eine rote Pille zuteilen und Patienten mit anderen Merkmalen eine grüne Pille, besteht das Ziel eines Angriffes auf dieses System beispielsweise darin, dass das System eine falsche Pille zuteilt und so einen Fehler macht.

Grosse illustrierte die Angriffe anhand von drei Wirkungskategorien: Die verbrauchten Ressourcen *(spent resources)*, die Beeinflussung der Leistung *(performance)* und die Übernahme oder Entwendung «geistigen Eigentums». Beim ersten erwähnten Angriffspunkt könne auch von *resource theft* gesprochen werden, da ein Angriff dazu füh-

1 A. Früh/D. Haux, Foundations of Artificial Intelligence and Machine Learning, Weizenbaum Series, 2022.

© 2024 sic! Stiftung, Bern / Helbing Lichtenhahn Verlag, Basel
Alle Rechte vorbehalten. Jede Verwertung in anderen als in den gesetzlich zugelassenen Fällen bedarf der schriftlichen Zustimmung des Verlages.
Tous droits réservés. Toute représentation ou reproduction, intégrale ou partielle, faite sans le consentement préalable de la maison d'édition, est interdite.
Auch auf www.legalis.ch und swisslex.ch / Également sur www.legalis.ch et swisslex.ch

ren könne, dass ein KI-System oder ein ML-Model in der Nutzung der eigenen Ressourcen (bspw. Energie, Zeit, Rechenleistung, etc.) beeinträchtigt oder ineffizienter werde. Ein Beispiel ist *adversarial initialization*, bei der bei einem neuronalen Netz die initialen Gewichte so verändert werden, dass das Training sich stark verlängert oder sogar ganz fehlschlägt.[2] Zur ersten Kategorie gehöre auch das *sponge poisoning*. Diese Art des Angriffs ermögliche es, den Stromverbrauch und die Latenzzeit von Modellen des maschinellen Lernens mit Eingaben zur Testzeit zufällig zu erhöhen.[3] Zur zweiten Gruppe der Angriffe auf die Leistung gehören gem. Grosse die *adversarial examples* oder *evasion*, *poisoning* und *backdooring*. *Adversarial examples* und *evasion* sind Eingaben, die absichtlich so gestaltet sind, dass sie ML-Modelle dazu verleiten, falsche Vorhersagen oder Klassifizierungen zu treffen *(adversarial examples)* oder widersprüchliche Beispiele zu liefern *(evasion)*. Während beim *poisoning* die Trainingsdaten oder deren Labels verändert werden, um die Genauigkeit des Klassifikators zu verringern, wird das häufigere *backdooring* als Auswahl von Eingabemustern beschrieben, die zuverlässig eine bestimmte falsche Klassifizierungsausgabe auslösen.[4] Zur dritten Kategorie gehörten gem. Grosse die vier Angriffe des *Model Reverse Engineering*, des *Model Stealing*, der *Membership Inference* und *Model Inversion*. Bei allen verlassen relevante Daten (z.B. Trainingsdaten oder das Modell selbst) die Sphäre des Anwenders des KI-Systems, ohne dass dieser dies möchte.

Die Referentin berichtete weiter aus ihrer eigenen Forschung, bei der das Ausmass von Angriffen auf KI-Systeme untersucht wurde. Dabei zeigte sich, dass von 139 Befragten etwa 16% Angriffe auf KI erlebt hatten. Von diesen bestand allerdings ein Drittel in «klassischen» *Cybersecurity*-Angriffen, bei denen das KI-System irrelevant war. Bei fünf Angriffen handelte es sich um *Poisoning*- und *Evasion*-Fälle und der Rest liess sich nicht zuordnen. Abschliessend zeigte sie auf, dass Angriffe in der Praxis oft ähnlich wie nicht-sicherheitsrelevante Fehler wahrgenommen werden. In anderen Worten wird in der Praxis oft angenommen, dass ein KI-System aus sich selbst heraus nicht funktioniert, seltener wegen eines Angriffes. Dies spiegelt ähnliche Wahrnehmungen in der Cybersicherheit von 1992 wieder.

2. Adversarial Attacks: (Potenzielle) Schäden

Dr. Dario Haux, Junior Associate in Zürich, referierte über tatsächliche und potenzielle Schäden, die ein Angriff auf ein KI-System zur Folge haben kann. Dabei unterschied der Referent zwischen Systemen der Text-, Bild-, Sprach- und Stimmerkennung.

Im Bereich der Bilderkennung, zeigte der Referent den Effekt von *adversarial noise*[5] auf. Ein KI-System könne durch eine nicht sichtbare Veränderung von Pixeln eines Bildes so beeinflusst werden, dass es nicht mehr in der Lage sei, ein Bild korrekt zu erkennen. Das KI-System könne dadurch auch dazu verleitet werden, einen falschen Bildinhalt wiederzugeben. Dies wurde mit dem Beispiel veranschaulicht, dass ein KI-System auf einem Bild eine Schüssel Guacamole wahrnahm, obwohl es sich für den menschlichen Betrachter offensichtlich um ein Bild einer grauen Katze handelte. Dies allein wäre an sich amüsant, jedoch zeigte der Referent schnell auf, weshalb dies zu grossen Schäden führen kann. Es wurde den Anwesenden ein Bild einer 3D-gedruckten Schildkröte gezeigt, welche aufgrund der gedruckten Musterung auf dem Schildkrötenpanzer vom KI-System als Gewehr identifiziert wurde. Der Referent machte darauf aufmerksam, dass es somit auch möglich wäre, dass ein KI-System eine Schusswaffe umgekehrt als Schildkröte wahrnimmt und somit wichtige Sicherheitsmassnahmen umgangen werden können.

In einem weiteren Beispiel zeigte der Referent zwei Bilder. Auf beiden Bildern war derselbe Mann abgebildet, wobei der einzige Unterschied darin bestand, dass er auf einem Bild eine Brille mit speziellen Mustern trug. Durch diese Brille wurde eine Gesichtserkennung-KI in die Irre geführt und statt dem Mann erkannte das KI-System das Gesicht von Jennifer Lopez. Mit diesem Beispiel betonte der Referent, dass durch Angriffe auf Bild- und Spracherkennungsprogramme Missbräuche im Zusammenhang mit Bankgeschäften möglich wären, etwa indem durch den gezielten Angriff auf die KI sowohl die Authentifizierung mittels Bild als auch jene mittels Stimme manipuliert werden könnte.

Als abschliessendes Beispiel zeigte der Referent noch den Effekt von *adversarial noise* auf ein Diagnostik-Tool auf. Ein Bild eines Muttermals wurde zusammen mit *adversarial noise* in ein Diagnostik-Tool eingespeist und das KI-System lieferte die Diagnose eines Tumors, obwohl es sich nicht um einen Tumor handelte.

Auf diese Weise zeigte Haux auf, dass es aufgrund der verschiedensten Arten der Angriffe auf KI-Systeme ebenfalls eine enorme Bandbreite potenzieller Schäden gibt.

3. Diskussion: Phänomen, Vielfalt, Gefahren, Handlungsbedarf

Nach den zwei Inputs von Grosse und Haux wurde die erste Diskussionsrunde eröffnet. Die Anwesenden diskutierten über das Phänomen, die immense Vielfalt und die daraus resultierenden Gefahren der Angriffe auf KI-Systeme und ML-Modelle. Im Rahmen dieser ersten Diskussionsrunde betonte Grosse, für die Bedeutung der Sicherheit an sich sei wichtig zu wissen, wie viele KI-Systemen und ML-Modelle[6]

2 K. Grosse/T. A. Trost/M. Mosbach/M. Backes/D. Klakow, On the security relevance of initial weights in deep neural networks, Artificial Neural Networks and Machine Learning, ICANN 2020, 13 ff.

3 A. Cinà/A. Demontis/B. Biggio/F. Roli/M. Pelillo, Energy-latency attacks via sponge poisoning, *arXiv preprint arXiv:2203.08147*, 2022.

4 Vgl. Fn. 3.

5 *Adversarial noise* ist eine gezielt eingefügte Störung, bspw. durch Veränderung von Pixel eines Bildes, um ein KI-System in die Irre zu führen und ein fehlerhaftes Ergebnis zu erwirken.

6 Die Referenten bezogen sich sowohl auf die Begriffe KI-System als auch auf ML-Model, wobei es sich beim ML-Model um eine spezifische Art von KI-System handelt. KI ist somit ein weit gefasster Begriff bzw. Überbegriff für maschinenbasierte Anwendungen, die die menschliche Intelligenz nachahmen. Nicht alle KI-Systeme sind ML-Modelle, aber sämtliche ML-Modelle sind KI-Systeme.

© 2024 sic! Stiftung, Bern / Helbing Lichtenhahn Verlag, Basel
Alle Rechte vorbehalten. Jede Verwertung in anderen als in den gesetzlich zugelassenen Fällen bedarf der schriftlichen Zustimmung des Verlages.
Tous droits réservés. Toute représentation ou reproduction, intégrale ou partielle, faite sans le consentement préalable de la maison d'édition, est interdite.
Auch auf www.legalis.ch und swisslex.ch / Également sur www.legalis.ch et swisslex.ch

angegriffen werden könnten. Wenn z.B. von 400 ML-Modellen zehn angegriffen werden könnten, dann seien vier Angriffe bereits viel und Sicherheitsfragen seien von grösster Bedeutung. Darüber hinaus waren sich die Diskutanten einig, dass das allgemeine Verständnis im Bereich der Entwicklung und genauen Funktionsweise der KI noch eher gering ist, insbesondere hinsichtlich ihrer Komplexität und ihres potenziellen Einsatzes. Daher sollte das Bewusstsein der Nutzer in dieser Hinsicht geschärft werden. Genannt wurden verschiedene Aspekte wie die soziale Verantwortung, die Haftung (selbst für kleine Angriffe wie z.B. die Änderung von Lebenslaufbewertungen für das Einstellungsverfahren), das Verständnis von Sicherheitsrisiken, die rechtliche Kontrolle über Daten, Fragen betreffend Bias sowie die Präzision und der Zweck des Einsatzes von KI. Es wurde darauf hingewiesen, dass der Zweck der Verwendung der KI die Risiken stark beeinflusse; es komme z.B. darauf an, ob man KI für private Zwecke oder zur Steuerung einer Rakete nutze.

III. Zweiter Teil: Technische und Rechtliche Massnahmen

1. Technische Massnahmen

Wie im ersten Teil wurden den Anwesenden zuerst die tatsächlich technisch möglichen Massnahmen nähergebracht. Dominic Piernot, Managing Director eines im IT Consulting tätigen Unternehmens, zeigte, dass KI unser Verständnis der Realität verändert hat und ernst genommen werden muss. Anhand von diversen Beispielen legte er dar, wie KI-Systeme in die Irre geführt werden können. In einem der Beispiele zeigte der Referent ein Bild eines STOP-Schildes, welches durch das Anbringen von kleinen Stickern so verändert wurde, dass ein Mensch das Schild weiterhin problemlos erkennen kann, das KI-System aber nicht. Trotz der Gefahr von Angriffen und der dargelegten Manipulationsmöglichkeiten vertrat *Piernot* die Meinung, KI werde immer mehr Vertrauen gewinnen (auch bei der Nutzung von ChatGPT am Arbeitsplatz), ähnlich wie es in der Vergangenheit beim Cloud-Computing der Fall gewesen sei.

Der Referent zeigte in seiner Präsentation auf, dass trotz der Risiken und Gefahren, die mit dem Einsatz von KI-Systemen verbunden sind, ein enormes Potenzial besteht, KI-Systeme für gute Zwecke einzusetzen. Dies unterstrich der Referent mit einer umfangreichen Liste positiver und negativer Anwendungen. Auf der einen Seite könnten KI-Systeme beispielsweise genutzt werden, um *Deep Fakes* zu erstellen. Auf der anderen Seite hingegen sei es KI-Systemen auch möglich, Betrugsversuche oder *Phishing*-Versuche zu erkennen.

Piernot schloss seinen Vortrag mit der klaren Aussage, trotz des Einsatzes von KI bestehe die Herausforderung darin, zu verstehen, was wahr und was gefälscht sei. Als mögliche Lösung nannte er die Verwendung von digitalen Wasserzeichen, mochte aber nicht prognostizieren, ob diese ausreicht.

2. Rechtliche Massnahmen: Strafrecht

Die strafrechtlichen Massnahmen zur allfälligen Sanktionierung von Angriffen auf KI-Systeme wurden von Colin Carter, Doktorand an der Juristischen Fakultät der Universität Basel, vorgestellt. Der Referent konzentrierte sich bei der Anwendbarkeit der strafrechtlichen Haftung auf Angriffe auf KI-Anwendungen im Strassenverkehr und in Bezug auf generative KI.

Bei Angriffen auf KI im Bereich des Strassenverkehrs, z.B. durch Manipulation eines STOP-Schildes, zeigte Carter, dass Tatbestände für die falsche bzw. veränderte Signalisierung und Markierung (Art. 98 SVG), Tötung und Körperverletzung (Art. 117, Art. 122 StGB etc.), oder Gefährdung des Lebens (Art. 129 StGB) einschlägig sein könnten.

Weiter zeigte er die mögliche strafrechtliche Haftung bei Angriffen auf Applikationen, die generative KI verwenden und thematisierte mögliche strafrechtliche Hürden bei der Herstellung generativer KI-Modelle. Er präsentierte hierfür drei Fallbeispiele. Im ersten Beispiel manipulierte jemand durch das Hinzufügen spezifischer Prompts das Verhalten eines E-Mail-Agenten so, dass dieser vertrauliche Daten bekannt gibt. Hier diskutierte der Referent die Anwendbarkeit der unbefugten Datenbeschaffung i.S.v. Art. 143 StGB sowie des unbefugten Eindringens in ein Datenverarbeitungssystem i.S.v. Art. 143bis Abs. 1 StGB. Bezüglich der Erstellung solcher Angriffe (ohne deren direkte Ausführung) zweifelt Carter an einer Strafbarkeit, denn Art. 143bis Abs. 2 StGB erfasst zwar «andere Daten», erfordert aber nach Abs. 1 auch einen besonderen Schutz gegen einen Zugriff und dass die Daten in Verkehr gebracht oder zugänglich gemacht werden. Im zweiten Beispiel beschrieb Carter die Situation, in welcher eine Person ihren Unterlagen spezifische Tokens hinzufügt, sodass jemand, der diese durch ein Sprachmodell laufen lässt, von diesem Sprachmodell beleidigt wird. Hier wurde wieder zwischen der Erfüllung der Tatbestände bezüglich der Ehrverletzung (Art. 173, Art. 174 und Art. 177 StGB) und der eigentlichen Herstellung des Angriffes unterschieden. Das Kriterium der Erkennbarkeit durch die betroffene Person oder Dritte ist massgebend, womit für die eigentliche Erstellung eine Lücke zu existieren scheint. Das letzte untersuchte Beispiel beschreibt eine Person X, die Prompt-Anleitungen mit spezifischen zusätzlichen Tokens veröffentlicht, die besonders künstlerische KI-erzeugte Bilder erstellen sollen. Werden die Prompts jedoch verwendet, generieren die Bildmodelle strafrechtlich relevante Pornografie und Gewaltdarstellungen. Im Bereich der strafrechtlich relevanten Pornografie (Art. 197 StGB) und Gewaltdarstellungen (Art. 135 StGB) ist bereits die Herstellung und der Besitz strafbar. Somit würden sich zum einen strafrechtliche Fragen auf der Seite des Nutzers des KI-Systems stellen, da dieser unweigerlich solche Inhalte generiert und schliesslich auch besitzt. Zum anderen stellen sich dieselben Fragen auch bei Person X, da diese bei der Erstellung des Angriffes unweigerlich testen musste, ob dieser funktioniert. In diesem Zusammenhang thematisierte der Referent auch die Schwierigkeit bei der Herstellung und

© 2024 sic! Stiftung, Bern / Helbing Lichtenhahn Verlag, Basel
Alle Rechte vorbehalten. Jede Verwertung in anderen als in den gesetzlich zugelassenen Fällen bedarf der schriftlichen Zustimmung des Verlages.
Tous droits réservés. Toute représentation ou reproduction, intégrale ou partielle, faite sans le consentement préalable de la maison d'édition, est interdite.
Auch auf www.legalis.ch und swisslex.ch / Également sur www.legalis.ch et swisslex.ch

dem Testen von generativen KI-Modellen, denn i.S. der Gefahrenverhinderung müssten die Hersteller eines KI-Systems sicherstellen, dass die Modelle nicht für die Generierung solcher Inhalte missbraucht werden können. Damit wäre aber möglicherweise auch die testende Person strafrechtlich belangbar, da auch hier die in einem Trainings- oder Testlauf generierten Inhalte die Straftatbestände erfüllen würden.

Zusammenfassend besteht nach der Ansicht Carters das strafrechtliche Haftungsrisiko bei der Herstellung von den umschriebenen Angriffen in wenigen speziellen Fällen. Es bestünden aber auch Fälle, wo die Herstellung solcher Angriffe in sich selbst noch zu keiner Strafbarkeit führen und erst durch die eigentliche Ausführung des Angriffes dann strafrechtlich erfasst sind. Er betonte, dass die strafrechtliche Haftung gestützt auf den Grundsatz *«nulla poena sine lege»* nicht ohne Weiteres ausgeweitet werden dürfe.

3. Rechtliche Massnahmen: Haftungsrecht

Die Fragen der zivilrechtlichen Haftung für Angriffe auf KI-Systeme wurden von Prof. Dr. Melinda Lohmann, Professorin für Wirtschaftsrecht an der Universität St. Gallen, behandelt. Lohmann präsentierte diverse Bereiche des Haftungsrechts, in welchen bereits Haftungsnormen bestehen, die potenziell auf Angriffe auf KI-Systeme anwendbar sind, und Bereiche, in welchen neue Normen möglich wären.

Als erstes erwähnte die Referentin das Beispiel der selbstfahrenden Fahrzeuge und die möglichen Haftungsgrundlagen für *adversarial attacks*. Das Strassenverkehrsgesetz sehe bereits in Art. 58 SVG eine Fahrzeughalterhaftung vor, die ebenfalls auf selbstfahrende Fahrzeuge anwendbar sei. Obwohl Hersteller versuchen sollten, möglichst viele Sicherheitsvorkehrungen gegen *adversarial attacks* zu treffen, sei es noch zu früh, zu sagen, wie sich das regulatorische Umfeld in diesem Gebiet gestalte.

In einem zweiten Teil ging die Referentin auf vertragliche und ausservertragliche Ansprüche bei Cyber-Attacken ein. Sie unterstrich die Tatsache, dass die Schweiz im Gegensatz zur EU (noch) kein horizontales Regelwerk speziell für KI entworfen habe. Eine technologieneutrale und vor allem innovationsfördernde Regulierung habe aber auch Vorteile. Nicht abschliessend geklärt sei hierzulande, ob es sich bei isolierter Software um ein Produkt handle und damit das Produkthaftungsrecht anwendbar sei. Bei in ein Endprodukt eingebetteter Software trete diese Problematik in den Hintergrund, weil das Endprodukt als Produkt qualifiziere. Im Produkthaftpflichtrecht gebe es bereits die Anforderung einer gewissen Sabotagefestigkeit. Am Beispiel der selbstfahrenden Fahrzeuge bedeute dies, dass sich diese nicht durch Attacken auf deren System (bspw. Signalerkennung) beeinflussen lassen dürften. In ihren Ausführungen kam Lohmann zum Schluss, die Systeme müssten im Entwicklungsprozess auf diese Angriffe getestet werden. Sollte hingegen eine Schwachstelle erst auftreten, nachdem das Produkt in Verkehr gebracht wurde, sei es die Pflicht des Produktherstellers, im Rahmen seiner Produktbeobachtungspflicht sofort tätig zu werden und den Fehler zu beheben.

Abschliessend wandte sich Lohmann der rechtlichen Situation in der EU zu. Die EU habe mit dem Entwurf des *AI-Acts* ein «Medley zwischen Produktsicherheit und Grundrechten» geschaffen. Jedoch enthalte dieser Rechtsakt keine Haftungsgrundlagen. Diese seien stattdessen in der *AI Liability Directive* und der *Product Liability Directive* enthalten. Die künftige *AI Liability Directive* solle nach derzeitigem Stand der Dinge eine ausservertragliche, verschuldensabhängige Haftung vorsehen. Das Verschulden könne in der Verletzung der im *AI-Act* normierten Sorgfaltspflicht gesehen werden, wenn beispielsweise die notwendige *cyber robustness* fehle. Die *Product Liability Directive* stehe im Gegensatz zu der *AI Liability Directive* kurz vor dem Inkrafttreten und sehe vor, dass eine Software als Produkt qualifiziert werde. Zudem seien die Anforderungen und Bestimmungen zur Fehlerhaftigkeit konkretisiert worden. Schliesslich wies die Referentin noch auf den *Cyber Resilience Act* hin, welcher die Sicherheit von Produkten mit digitalen Elementen fördern und Sicherheitsmassnahmen vorsehen solle. Zusammenfassend bleibe jedoch auch hier *«it's too soon to tell»* die zentrale Aussage, wenn es um haftungsrechtliche Massnahmen in Bezug auf Angriffe auf KI-Systeme gehe. Zudem müsse man abwarten, wie sich die KI-Systeme weiterentwickelten.

4. Rechtliche Massnahmen: Immaterialgüterrecht

In der letzten Präsentation ging Prof. Früh der Frage nach, ob das Immaterialgüterrecht zum Schutz vor Angriffen auf KI-Systeme eingesetzt werden könne. Dies hänge einerseits davon ab, ob gewisse Teile des Systems oder das System als Ganzes immaterialgüterrechtlich geschützt seien. Andererseits müsste es sich bei den jeweiligen Angriffen um Nutzungen handeln, die dem jeweiligen Rechteinhaber oder der Rechteinhaberin vorbehalten seien. Zudem müsse die jeweilige Nutzung weder gesetzlich (als Schranke) noch vertraglich (in Form einer Lizenz) erlaubt sein.

Der Referent stellte die möglichen Gegenstände des immaterialgüterrechtlichen Schutzes vor und würdigte kursorisch deren potenzielle Schutzfähigkeit. Dabei seien fünf verschiedene Elemente zu unterscheiden: *Erstens* könnten KI-Frameworks, die den Betrieb des KI-Systems unterstützen oder ermöglichen (wie z.B. TensorFlow oder PyTorch) durch (Software-)Urheberrechte geschützt sein. *Zweitens* sei die Architektur, also der konkrete Aufbau eines KI-Systems (wie z.B. ein *Deep Neural Network*, eine *Support Vector Machine* oder ein *Random Tree Forest*), ein mögliches Schutzobjekt. Ein Schutz sei aber hierfür weder unter patentrechtlichen noch unter urheberrechtlichen Gesichtspunkten wahrscheinlich. Dies sei, *drittens*, auch bei der Methode des Lernens (wie z.B. *supervised*, *unsupervised* oder *reinforcement learning*) der Fall, zumal nach Art. 52 EPÜ Algorithmen als solche vom Patentschutz ausgenommen sind und auch nicht urheberrechtlich geschützt werden können. *Viertens* bestünden an den Daten (namentlich den Trainingsdaten) ebenfalls keine Rechte, jedenfalls dann, wenn es um den Schutz des gesamten Datenbestandes gehe und nicht um den Schutz einzelner Daten wie Bilder oder Texte. *Fünftens*

© 2024 sic! Stiftung, Bern / Helbing Lichtenhahn Verlag, Basel
Alle Rechte vorbehalten. Jede Verwertung in anderen als in den gesetzlich zugelassenen Fällen bedarf der schriftlichen Zustimmung des Verlages.
Tous droits réservés. Toute représentation ou reproduction, intégrale ou partielle, faite sans le consentement préalable de la maison d'édition, est interdite.
Auch auf www.legalis.ch und swisslex.ch / Également sur www.legalis.ch et swisslex.ch

sei auch der Schutz des trainierten Modells zweifelhaft. Immerhin sei dieses möglicherweise als computerimplementierte Erfindung vom Patentrecht oder vom Urheberrecht geschützt, sofern es die Form eines Computerprogramms aufweise.

Eine Verletzungshandlung, führte Früh weiter aus, liege im Urheberrecht dann vor, wenn es zu einer Vervielfältigung, Verbreitung oder öffentlichen Wiedergabe (bzw. Wahrnehmbarmachung) komme. In Bezug auf KI-Angriffe könnte der Rechteinhaber folglich nur gegen einen Angreifer vorgehen, wenn dieser ein Modell extrahiere und dieses als Computerprogramm verwende – z.B. auch zur Herstellung von *adversarial examples*. Wäre ein trainiertes Modell hingegen als computerimplementierte Erfindung geschützt, wäre dieser Schutz durch die meisten Angriffe typischerweise verletzt.

Früh betonte, dass das Immaterialgüterrecht trotz eines möglichen punktuellen Einsatzes kein kohärentes Instrument gegen Angriffe auf KI-Systeme biete. Zudem dürfe eine Mobilisierung des Immaterialgüterrechts gegen Angriffe auf KI-Systeme nicht zu einer Ausweitung der Funktionen des Immaterialgüterrechtsschutzes (Innovations- und Kreationsschutz) führen.

5. Diskussion: Wirksamkeit Massnahmen de lege lata, Differenzierungen, Regelungsbedarf, Regelungsumfang, Regelungsebene

Die abschliessende Diskussion erfasste eine breite Palette zentraler, von den Referentinnen und Referenten vorgebrachter Themen. Aus der Diskussion ging hervor, dass die Qualität der KI-Systeme stark variieren kann. Entsprechend ist damit zu rechnen, dass auch unsichere Systeme eingesetzt werden oder auf den Markt kommen. Dies warf die Frage nach Qualitätssicherungssystemen und Compliance-Systemen in Bezug auf KI-Produkte auf.

Gleichzeitig rückten so auch Fragen der Haftung und Verantwortung in den Mittelpunkt der Diskussion. Schwierigkeiten sind die oft diskutierte Intransparenz der Systeme aber auch die Tatsache, dass einzelne Komponenten von verschiedenen Herstellern stammen können. Entsprechend ergeben sich Probleme bei der eindeutigen Zuordnung der Verantwortung. In diesem Zusammenhang wies Carter auf die Möglichkeit hin, dem Hersteller eines KI-Systems eine Garantenstellung im Sinne von Art. 11 Abs. 2 StGB zuzuschreiben.

Weiter wurde in der Diskussion die Wichtigkeit der Datenqualität hervorgehoben. Vor allem Grosse erinnerte daran, dass Angriffe, welche die Trainingsdaten verändern, sich negativ auf die Datenqualität auswirken. Qualitativ und quantitativ nutzbare Daten seien ohnehin schwer zu bekommen und Angriffe könnten diese wichtige Ressource zusätzlich beeinträchtigen.

Kontrovers diskutiert wurden auch ökonomische Aspekte. So wurde von Workshop-Teilnehmern darauf hingewiesen, dass das Training von *Large Language Models* (LLM) mit immensen Kosten verbunden sei, was zwei Konsequenzen habe: Einerseits befänden sich die LLM deshalb in den Händen weniger grosser Akteure. Kleinere Marktteilnehmer, wie z.B. Start-Ups, seien auf die grossen Akteure angewiesen, und zwangsläufig der Gefahr ausgeliefert, dass ihre eigenen Resultate durch die grossen Akteure verfälscht werden könnten. Entsprechend wurde z.T. vertreten, die Haftung müsse davon abhängig gemacht werden, ob es sich beim KI-System um ein Open-Source Modell handle oder nicht. Andererseits stelle sich die Frage, ob die trainierten Modelle mangels Rechtsschutzes ausreichend vor einer Übernahme durch Dritte geschützt seien. In diesem Zusammenhang wurde angemerkt, naheliegender als eine weitere Ausdehnung des patent- oder urheberrechtlichen Schutzes sei ein Leistungsschutzrecht. Gegen ein solches spreche umgekehrt die Tatsache, dass es momentan weder an proprietären noch an offenen LLMs fehle und der faktische Schutz dieser Systeme offensichtlich ausreiche.

IV. Zusammenfassung

Insgesamt boten die Vorträge und die Diskussionen einen tiefen Einblick in die technischen und rechtlichen Herausforderungen, die sich mit der fortschreitenden Entwicklung von KI und ML ergeben. Die kurzen Inputs ermöglichten eine rasche Orientierung und liessen ausreichend Raum für die interdisziplinären Diskussionen. Dabei zeigte sich, dass es kein Patentrezept zum Schutz vor Angriffen auf KI-Systeme gibt. Eine Suche nach rechtlichen Lösungen gegen solche Angriffe berührt ganz offensichtlich grundlegende Fragen: Jene nach den strafrechtlich oder haftungsrechtlich geschützten Rechtsgütern sowie jene nach der Funktion der Immaterialgüterrechte.

Der Austausch führte allen Teilnehmerinnen und Teilnehmern einmal mehr vor Augen, wie rasch die Entwicklung von KI-Systemen voranschreitet. Im gleichen Tempo muss sich indes auch die Diskussion um die Sicherheit und den Missbrauch dieser Systeme weiterentwickeln. Der Tech Law Workshop war hierzu ein wichtiger erster Schritt.[7]

7 Zeitnahe zu diesem Tagungsbericht erfolgt die Veröffentlichung des folgenden Beitrags: A. Früh/D. Haux, Countermeasures against Adversarial Attacks on Computational Law, The Journal of Cross-Disciplinary Research in Computational Law. Der Beitrag fokussiert sich u.a. auf die potenziellen rechtlichen Massnahmen, um sich gegen Adversarial Attacks zu verteidigen.

© 2024 sic! Stiftung, Bern / Helbing Lichtenhahn Verlag, Basel
Alle Rechte vorbehalten. Jede Verwertung in anderen als in den gesetzlich zugelassenen Fällen bedarf der schriftlichen Zustimmung des Verlages.
Tous droits réservés. Toute représentation ou reproduction, intégrale ou partielle, faite sans le consentement préalable de la maison d'édition, est interdite.
Auch auf www.legalis.ch und swisslex.ch / Également sur www.legalis.ch et swisslex.ch