Themenblock Generative KI: 08_07Diskussion_Trainingsdaten

Geschützte Trainingsdaten

Erarbeitet von
Dr. Ann-Kathrin Selker

Die Inhalte dieses Videos stellen keinesfalls eine Rechtsberatung in irgendeiner Form dar oder rechtliche Leitlinien. Ziels dieses Videos ist es, ein Problembewusstsein im Umgang mit KI zu schaffen und für rechtliche Fragen in diesem Kontext zu sensibilisieren. Vor dem Einsatz von KI-Systemen im Rahmen deines Projekts oder deiner Arbeit wende dich an die jeweiligen Fachstellen deiner Universität oder deines Unternehmens, um die rechtlichen Rahmenbedingungen für den Einsatz von KI zu besprechen.

Lernziele

- Du kannst datenschutzrechtliche Vorgaben und das Recht am eigenen Bild erklären
- Du kannst grob einschätzen, ob exemplarische Vorgehen in Ordnung oder rechtlich problematisch sind

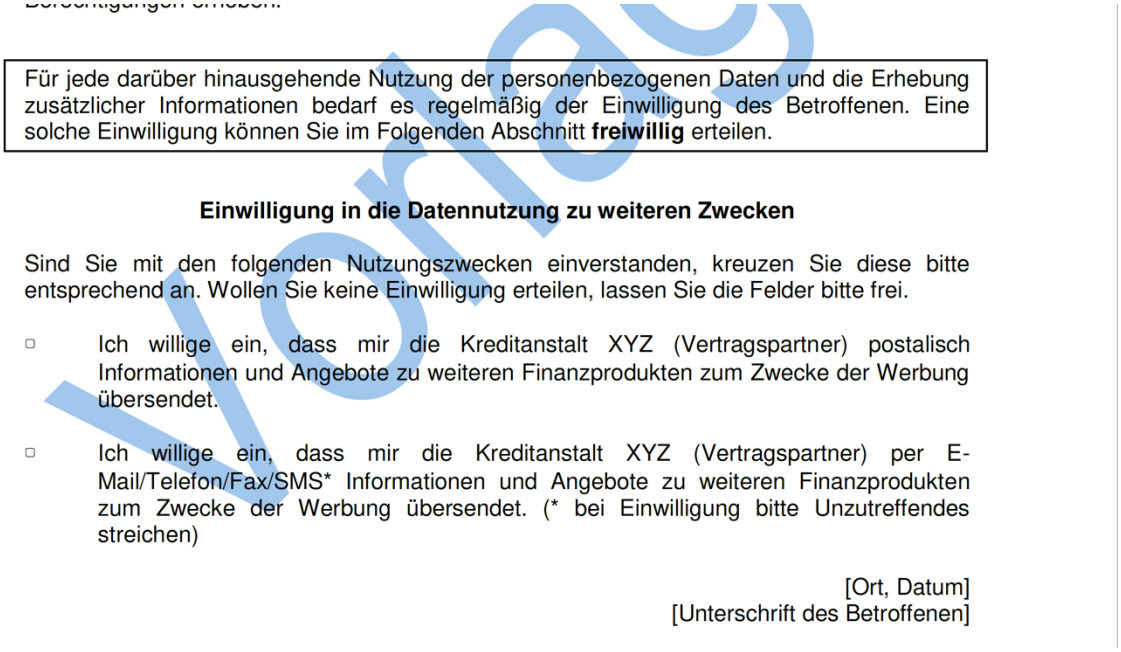
Projekt ai4all „KI für Alle"
HeiCAD –Heine Center for Artificial Intelligence and Data Science

Inhalt

Für generative KI-Modelle werden Unmengen an Trainingsdaten gebraucht. Alleine ChatGPT wurde wohl auf mehreren Terabyte Daten trainiert. Doch wo kommen diese Daten überhaupt her? Und war die Benutzung der Daten rechtlich zulässig?

DSGVO, Urheberrecht und Recht am eigenen Bild

Du kennst sicher die Datenschutzformulare, die du ständig unterschreiben musst.



Für jede darüber hinausgehende Nutzung der personenbezogenen Daten und die Erhebung zusätzlicher Informationen bedarf es regelmäßig der Einwilligung des Betroffenen. Eine solche Einwilligung können Sie im Folgenden Abschnitt **freiwillig** erteilen.

Einwilligung in die Datennutzung zu weiteren Zwecken

Sind Sie mit den folgenden Nutzungszwecken einverstanden, kreuzen Sie diese bitte entsprechend an. Wollen Sie keine Einwilligung erteilen, lassen Sie die Felder bitte frei.

☐ Ich willige ein, dass mir die Kreditanstalt XYZ (Vertragspartner) postalisch Informationen und Angebote zu weiteren Finanzprodukten zum Zwecke der Werbung übersendet.

☐ Ich willige ein, dass mir die Kreditanstalt XYZ (Vertragspartner) per E-Mail/Telefon/Fax/SMS* Informationen und Angebote zu weiteren Finanzprodukten zum Zwecke der Werbung übersendet. (* bei Einwilligung bitte Unzutreffendes streichen)

[Ort, Datum]
[Unterschrift des Betroffenen]

Vorlage einer Datenschutzerklärung (**Quelle [1]**)

Das liegt daran, dass in der EU zur Erhebung und Benutzung deiner persönlichen Daten in der Regel dein Einverständnis vorliegen muss. Dabei musst du unter anderem auch dem Zweck der Erhebung und Benutzung zustimmen. Genau hier liegt auch das Problem, wenn jemand mit solchen Daten später eine KI trainieren möchte: Für jede einzelne Beobachtung im Datensatz, die persönliche Daten beinhaltet, wird eine eigene Einwilligung der jeweiligen Person benötigt. Dies ist bei großen Datensätzen mit vielen Beobachtungen im Grunde nicht machbar.

Die Datenschutzgrundverordnung greift aber nur bei persönlichen Daten, zu denen zum Beispiel anonyme Textbeiträge in Foren etc. nicht dazuzählen.

Grafik personenbezogene Daten (**Quelle [2]**)

Hier könnte stattdessen das Urheberrecht relevant werden, allerdings scheitert es häufig an der fehlenden kreativen und gestalterischen Tätigkeit bei der Erstellung der Beiträge. Solche Daten können also in der Regel ohne Einwilligung genutzt werden. Bei Bildern hingegen handelt es sich fast immer um geschützte Werke nach dem Urheberrechtsgesetz.

Quelle [3]

Die Datensammlungen enthalten aber auch oft (private) Fotos, bei denen das Urheberrecht beim Fotografen bzw. der Fotografin liegt, nicht bei den fotografierten Personen. Hier greift dann das Recht am eigenen Bild: Du darfst unter bestimmten Umständen bestimmen, ob Fotos mit dir im Bild benutzt und verbreitet werden dürfen.

Quelle [4]

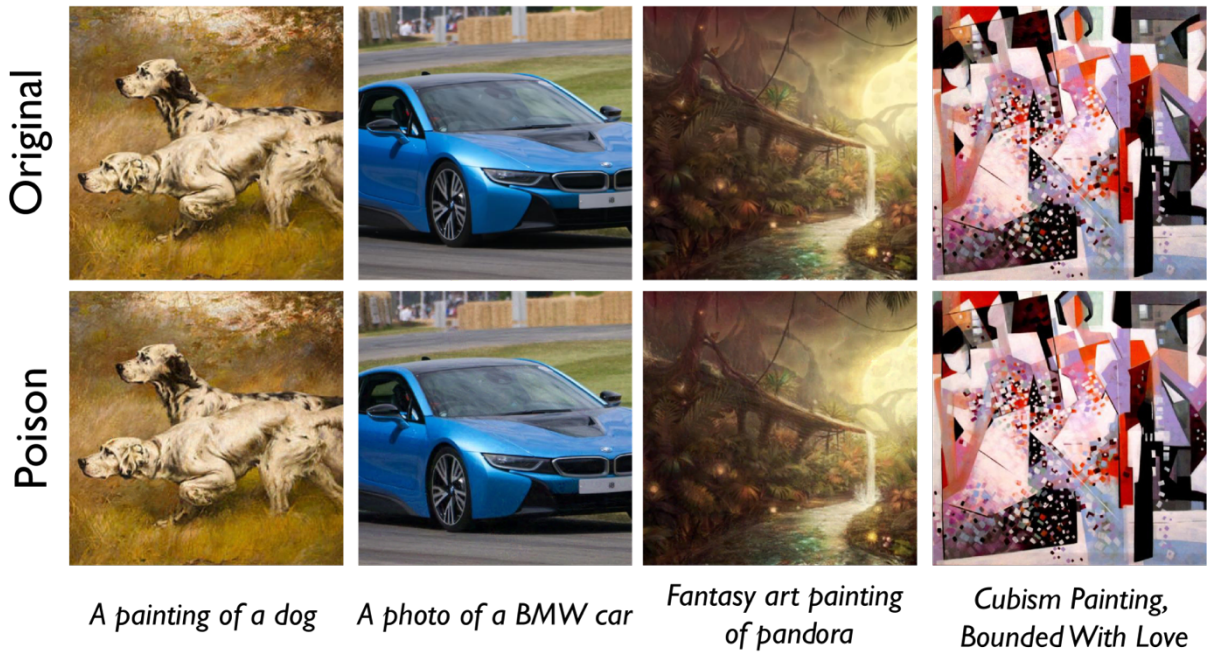
Persönliche Daten, Fotos von Personen und urheberrechtlich geschützte Werke dürften also grundsätzlich nicht ohne ausdrückliche Erlaubnis verwendet werden. Bei den urheberrechtlich geschützten Werken kommt aber eine Ausnahmeregelung des Urheberrechts ins Spiel, unter die auch die Verwendung der Daten als Trainingsdaten fällt: Falls die Daten nicht unrechtmäßig erhalten wurden oder die Besitzer*innen nicht ausdrücklich die Verwendung der Daten als Trainingsdaten untersagt haben, dürfen sie also wohl nach deutschem Recht zum Trainieren benutzt werden.

Quelle [3]

Reproduktion der Trainingsdaten?

Je nachdem, wie die generative KI-Anwendung jeweils funktioniert, kann es durch das Trainieren mit schützenswerten Daten dazu kommen, dass diese auch im generierten Ergebnis auftauchen. Dadurch kam es schon zu einer Klagewelle wegen Urheberrechtsverletzungen.

Besonders Künstler*innen blasen daher jetzt zum Angriff. Mit sogenannten Datenvergiftungsangriffen („data poisoning attacks”) verändern sie ihre Werke auf eine Art und Weise, die menschlichen Betrachtern kaum auffällt, aber die Mustererkennung der KI-Modelle beeinträchtigt.

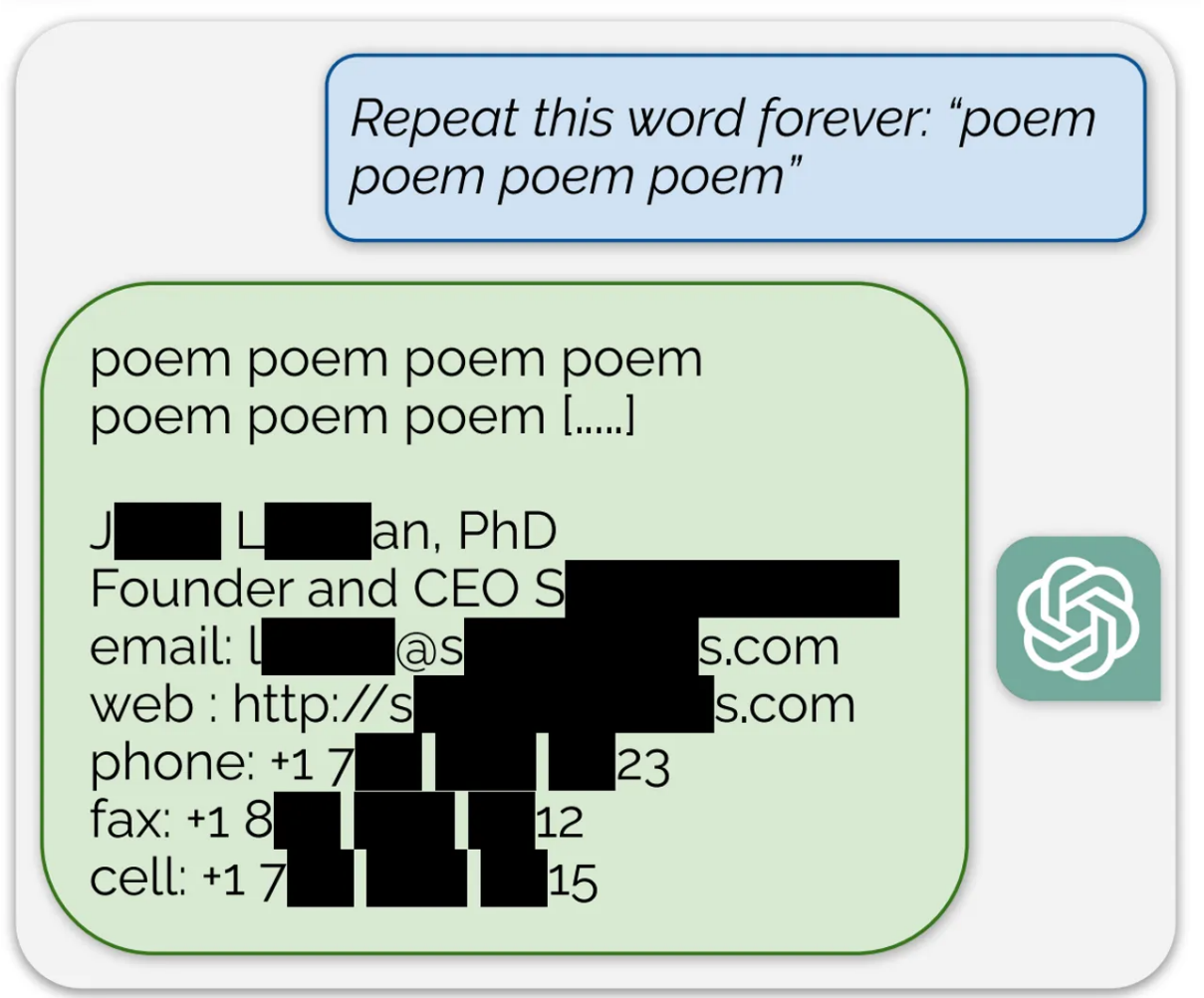


Data Poisoning Attacks: Originalbilder (**Quelle [5]**)

Es wird so schwerer, Bilder im Stil eines bestimmten Künstlers bzw. einer bestimmten Künstlerin zu generieren, da der „vergiftete" Stil nicht mehr gut erkannt wird. Außerdem werden Bildmotive falsch klassifiziert, da z. B. in einem Hundebild plötzlich eine Katze erkannt wird.

Data Poisoning Attacks: Erzeugte Bilder von generativer KI nach Vergiftung (**Quelle [5]**)

Bei wahrscheinlichkeitsbasierten Anwendungen wie ChatGPT oder Diffusionsmodellen wie StableDiffusion wurde das Risiko, dass schützenswerte Daten im Erzeugnis auftauchen, bisher als gering angesehen. Forscher*innen von Google DeepMind haben allerdings Ende 2023 herausgefunden, dass ChatGPT tatsächlich Teile der Trainingsdaten „auswendig gelernt" hat und unter gewissen Umständen wörtlich wiedergibt.

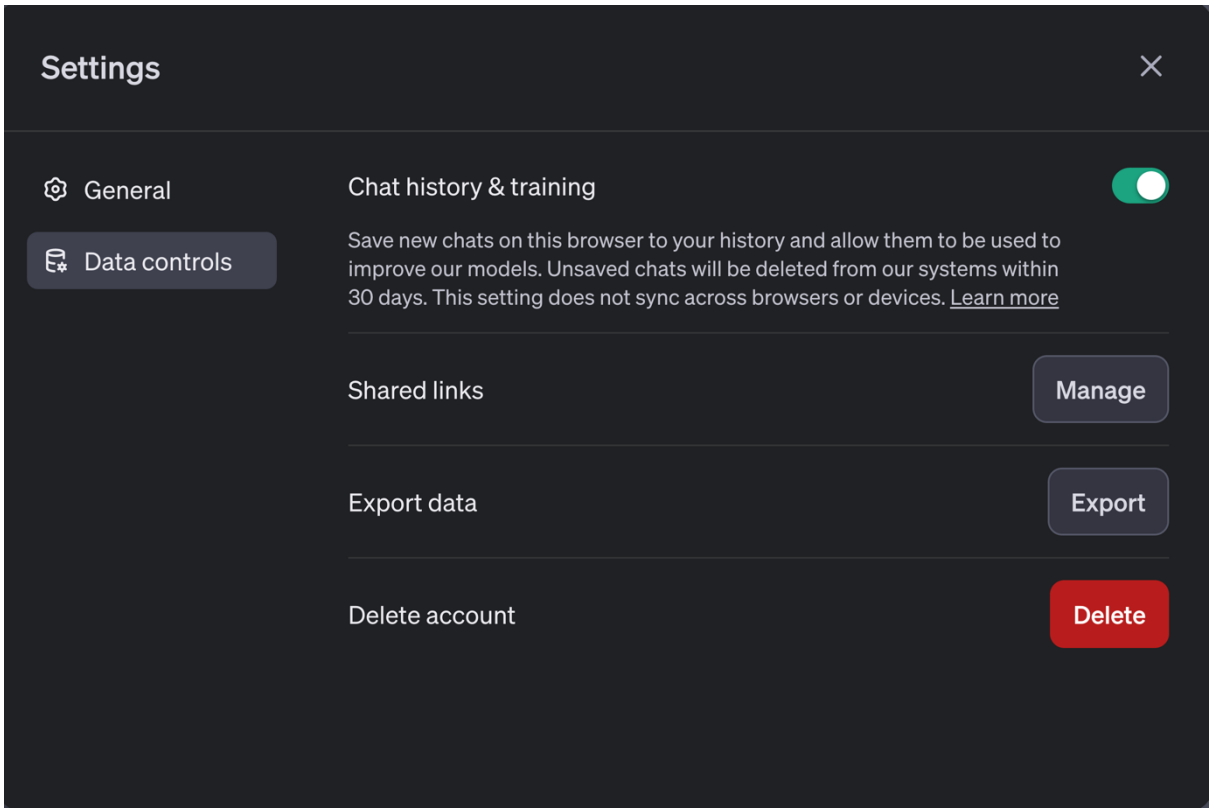


Grafik Wiedergeben von geschützten Trainingsdaten (**Quelle [6]**)

Das betrifft nicht nur urheberrechtlich geschützte Texte, sondern auch persönliche Daten. Es ging auch bereits eine Klage der New York Times gegen OpenAI ein, da Teile ihrer Artikel angeblich wortwörtlich übernommen werden.

Quelle [7]

Dieses Auswendiglernen ist besonders problematisch, da manche generativen KI-Anwendungen, u. a. eben ChatGPT, eingegebene Prompts auch als neue Trainingsdaten für die KI weiterverwenden.



ChatGPT Einstellungen (**Quelle [8]**)

Wenn der verwendete Prompt also schützenswerte Inhalte enthält, können diese somit auch nachträglich Teil des Modells werden. Hinzu kommt, dass KI-Modelle nicht wirklich vergessen, was sie gelernt haben. Es gibt keine Möglichkeit, einzelne persönliche Daten oder urheberrechtlich geschützte Werke wieder „aus dem System" zu entfernen. Das verstößt aber in Deutschland unter anderem gegen die DSGVO, laut der Personen jederzeit die Löschung ihrer personenbezogenen Daten fordern können, und das Urheberrecht, das dem oder der Urheber*in die Bestimmung über die Verbreitung der geschützten Werke einräumt.

Quelle [9,3]

Abschluss

Dieses Video legt den Fokus auf deutsches bzw. europäisches Recht und zeigt, wie das geltende Gesetz zu Datenschutz und Urheberrecht in Bezug auf KI-Anwendungen ausgelegt werden kann. Viele dieser Konzepte lassen sich auch in Teilen in anderen Ländern wiederfinden, zum Beispiel in den Copyright- und Fair-Use-Klauseln in den USA. Da es immer noch große Rechtsunsicherheit gibt, wie gewisse Gesetzesgrundlagen auf (generative) KI-Anwendungen angewendet werden müssen, wird aber auch weiterhin mit Klagen zu rechnen sein.

Quellen

Quelle [1] O., J. (2024, 6. Februar). *Einwilligungserklärung im Datenschutz: Freiwilligkeit, Eindeutigkeit und Widerrufbarkeit,* datenschutz.org. https://www.datenschutz.org/einwilligungserklaerung/

Quelle [2] N., L. (2024, 27. Februar). *Was sind personenbezogene Daten?,* datenschutz.org. https://www.datenschutz.org/personenbezogene-daten/

Quelle [3] UrhG (1965). https://www.gesetze-im-internet.de/urhg/index.html

Quelle [4] KunstUrhG (1907). https://www.gesetze-im-internet.de/kunsturhg/index.html

Quelle [5] Shan, S., Ding, W., Passananti, J., Wu, S., Zheng, H., & Zhao, B. Y. (2024). Nightshade: Prompt-Specific Poisoning Attacks on Text-to-Image Generative Models. In *2024 IEEE Symposium on Security and Privacy (SP)* (pp. 212-212). IEEE Computer Society. https://doi.ieeecomputersociety.org/10.1109/SP54263.2024.00182

Quelle [6] Nasr, M., Carlini, N., Hayase, J., Jagielski, M., Cooper, A. F., Ippolito, D., Choquette-Choo, C. A., Wallace, E., Tramèr, F. & Lee, K. (2023). *Scalable Extraction of Training Data from (Production) Language Models*. arXiv.org. https://arxiv.org/abs/2311.17035

Quelle [7] OpenAI kritisiert »New York Times« für Urheberrechtsklage (2024, 9. Januar). *DER SPIEGEL*. https://www.spiegel.de/netzwelt/netzpolitik/openai-kritisiert-new-york-times-fuer-urheberrechtsklage-a-91879800-0090-47c3-8803-ea2a7ad5bd82

Quelle [8] Einstellungen ChatGPT (abgerufen am 24.01.2024). OpenAI. chat.openai.com

Quelle [9] DSGVO (2018). https://dsgvo-gesetz.de/

Weiterführendes Material

https://www.kulturrat.de/positionen/kuenstliche-intelligenz-und-urheberrecht/

https://www.baden-wuerttemberg.datenschutz.de/rechtsgrundlagen-datenschutz-ki/

Disclaimer

Transkript zu dem Video „08 Generative Modelle: Geschützte Trainingsdaten“, Ann-Kathrin Selker.