



Wer AI sagt, muss auch Governance sagen

Verantwortung & Vertrauen in einer algorithmusgetriebenen Welt

Vienna Data Science Group

- **Mission**
 - Nonprofit association promoting knowledge, educating people and caring about social impact of data science methods and AI applications. We build a strong community of data scientists in Europe and all around the world. Our members are passionate data scientists from various areas of research and industry.
- **Diverse members**
 - Academics, professionals, students and all other Data Science enthusiasts
- **Different fields**
 - Mathematics, physics, econometrics, electrical engineering, medical science, finance, real estate, computer science and social sciences
- **World wide scope**

<https://viennadatasciencegroup.at/>

<https://viennadatasciencegroup.at/data4good/>

Künstliche Intelligenz ist ...

... wenn ein Algorithmus Aufgaben erledigt, die wir normalerweise einer menschlichen Intelligenz zuordnen (Text verstehen, Bilder erkennen, Entscheidungen treffen)

Wenn der Algorithmus...

1. ... aus Schritt für Schritt Anweisungen besteht
 - nennt man das symbolic AI
2. ... anhand von vielen Beispielen lernt
 - nennt man das Maschinelles Lernen
 - Deep Learning ist nur eines von vielen möglichen Methoden

Künstliche Intelligenz ist ...

... weder objektiv, noch neutral

Welche Anwendungen werden entwickelt:

- Vincent, J., *Deepfake bots on Telegram make the work of creating fake nudes dangerously easy*, the Verge, Oktober 2020, <https://www.theverge.com/2020/10/20/21519322/deepfake-fake-nudes-telegram-bot-deepnude-sensity-report>

Welche Funktionalitäten sollen sie haben:

- Duhaime-Ross, A., *Apple promised an expansive health app, so why can't I track menstruation?*, the Verge, September 2014, <https://www.theverge.com/2014/9/25/6844021/apple-promised-an-expansive-health-app-so-why-cant-i-track>

Wie und was wird klassifiziert:

- Leufer, D., *Computers are binary, people are not: how AI systems undermine LGBTQ identity*, Access Now, April 2021, <https://www.accessnow.org/how-ai-systems-undermine-lgbtq-identity/>

Diskriminierung, Vorurteile, Bias ...

- Simonite, T., ***When It Comes to Gorillas, Google Photos Remains Blind***, Wired, November, 2018 <https://www.wired.com/story/when-it-comes-to-gorillas-google-photos-remains-blind/>
- Simonite, T., ***When AI Sees a Man, It Thinks 'Official.' A Woman? 'Smile'***, November 2020, <https://www.wired.com/story/ai-sees-man-thinks-official-woman-smile/>
- The Guardian, ***Amazon ditched AI recruiting tool that favored men for technical jobs***, Reuters, Oktober, 2018 <https://www.theguardian.com/technology/2018/oct/10/amazon-hiring-ai-gender-bias-recruiting-engine>

Künstliche Intelligenz ist ...

... weder objektiv, noch neutral!

Welche Anwendungen werden ...

- Vincent, J., *Deepfake*,
the Verge, [Circulating](#)
[nudes-teleg](#)

*is dangerously easy,
deepfake-fake-*

... cherchez l'homme

Welche Funktionen ...

- Duhaime-Ross, *AI*,
the Verge, Septe
[expansive-health](#)

*menstruation?,
re-promised-an-*

Wie und was wird klas ...

- Leufer, D., *Computer*,
Access Now, April 20

*... now AI systems undermine LGBTQ identity,
[https://www.accessnow.org/how-ai-systems-undermine-lgbtq-identity/](#)*

KI Ethik Richtlinien

- **OECD** <https://www.oecd.ai/ai-principles>
- **UNESCO:** <https://en.unesco.org/news/unesco-launches-worldwide-online-public-consultation-ethics-artificial-intelligence>
- **UNICEF:** <https://www.unicef.org/globalinsight/reports/policy-guidance-ai-children>
- **EU:** https://ec.europa.eu/info/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en
- **Microsoft AETHER:** <https://www.microsoft.com/en-us/ai/responsible-ai>
- **Google's** (former) Ethics Committee: <https://www.reuters.com/article/us-alphabet-google-ai-idUSKCN1RH00S>
- **Partnership on AI:** <https://www.partnershiponai.org/>

Mit Stand 2019*, gibt es über 84 KI Ethik Richtlinien.

Entwurf einer EU KI Regulierung

Article 10: Data and Data Governance

Datensätze von hoch-risiko KI Anwendungen sollen folgende Anforderungen erfüllen:

- Qualitativ hoch-wertig (relevant, repräsentativ, komplett, fehlerfrei)
- Training und Testdaten unterliegen einem Data Governance und Management Regime, welches den Datenverwertungszyklus, angefangen mit der Datensammlung, über Pre-processing, Labelling, bis hin zum Einsatz in der Modellierung und Modellbewertung, nachvollziehbar und transparent macht.
- In der vorgesehenen technischen Dokumentation sollen auch Datenblätter beinhaltet sein, die die eingesetzten Daten ausführlich beschreiben. (**Annex IV**)

Datasheets for Datasets

In the electronics industry, every component, no matter how simple or complex, is accompanied with a datasheet that describes its operating characteristics, test results, recommended uses, and other information. By analogy, we propose that every dataset be accompanied with a datasheet...

In der Elektronikindustrie wird jede Komponente, egal wie einfach oder komplex, von einem Datenblatt (Datasheet) begleitet, dass die Betriebseigenschaften, Testergebnisse, empfohlene Anwendungen, und weitere Informationen enthält. Wir schlagen analog vor, dass jeder Datensatz von einem Datenblatt begleitet wird ...

Gebru, T. et al., *Datasheets for Datasets*, <http://arxiv.org/abs/1803.09010>

Bender, E. et al, *Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science*. Transactions of the Association for Computational Linguistics 6 (2018), 587–604.

Datasheets for Datasets

Wozu?

- Der Datensatz wird oft von einer anderen Gruppe erstellt, als die, die ihn anwendet
- Datensatz ErstellerInnen sollen sich überlegen, welche Daten sie zur Problemlösung brauchen
- Mehr Information für Regierungen, Zivilgesellschaft, und Endbenutzer

Vorteile:

- Erhöhte Transparenz und Verantwortlichkeit bei den Datensatzerstellern
- Resultate sind für andere Forscher leichter nachzuvollziehen
- Ungewollte Biases werden früher erfasst und können vermindert werden
- Anwender können besser beurteilen, welcher Datensatz für ihr Problem das Geeignetste ist

Jeder Datensatz hat seine eigene Geschichte

In einem Datenblatt soll festgehalten werden:

- Warum, für wen, und zu welchem Zweck wurde der Datensatz erstellt
- Woraus besteht der Datensatz
- Wie wurde er gesammelt/zusammengestellt/gelabelt
- Was waren die Pre-processing/Bereinigungs-Schritte
- Welche Anwendungen sind für den Datensatz vorgesehen
- Wie wird der Datensatz verteilt
- Wer ist für die Instandhaltung zuständig

Datasheets for Datasets

Datasheet Fragen:

- Wer war am Datensammelprozess beteiligt (z.B. Studenten, Crowdworkers, Contractors), und wie wurden sie für ihre Arbeit entlohnt (z.B. was haben die Crowdworkers verdient)
- Wurden die Daten gelabelt? ... Falls ja, bitte beschreiben.

Samuels, A., ***The Internet Is Enabling a New Kind of Poorly Paid Hell***, in The Atlantic, January 23 (2018)
<https://www.theatlantic.com/business/archive/2018/01/amazon-mechanical-turk/551192/>

Ghaffary, S., **The algorithms that detect hate speech online are biased against black people**, Vox, Aug 2019,
<https://www.vox.com/recode/2019/8/15/20806384/social-media-hate-speech-bias-black-african-american-facebook-twitter>

Simonite, T., ***When AI Sees a Man, It Thinks 'Official.' A Woman? 'Smile'***, November 2020,
<https://www.wired.com/story/ai-sees-man-thinks-official-woman-smile/>

Datasheets for Datasets

Datasheet Fragen:

- Wenn der Datensatz ein Sample aus einer größeren Menge ist, was war die Sampling-Strategie (z.B. deterministisch, probabilistisch, usw.)
- Wenn der Datensatz ein Sample aus einer größeren Menge ist, was ist die größere Menge? Ist der Sample repräsentativ? Falls ja, wie wurde das verifiziert? Falls nein, wieso nicht?
- Werden im Datensatz irgendwelche demographischen Untermengen identifiziert (z.B. Alter, Herkunft, Gender)? Falls ja, bitte beschreiben, wie diese identifiziert werden, und die jeweilige Aufteilung angeben.

Buolamwini, J. and Gebru, T., ***Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification***, Proceedings of Machine Learning Research 81:1–15 (2018) <http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf>

Simonite, T., ***When It Comes to Gorillas, Google Photos Remains Blind***, Wired, November 1, 2018 <https://www.wired.com/story/when-it-comes-to-gorillas-google-photos-remains-blind/>

The Guardian, ***Amazon ditched AI recruiting tool that favored men for technical jobs***, Reuters, Oktober, 2018 <https://www.theguardian.com/technology/2018/oct/10/amazon-hiring-ai-gender-bias-recruiting-engine>

O'Neil, C., ***Mutant Algorithms Are Coming for Your Education, Bloomberg***, September 2020, <https://www.bloomberg.com/opinion/articles/2020-09-08/mutant-algorithms-are-coming-for-your-education>

Datasheets for Datasets ... oder auch für Algorithmen?

Was ist die Optimierungsfunktion des eingesetzten Algorithmus?

- Ziad Obermeyer et al. ***Dissecting racial bias in an algorithm used to manage the health of populations***. In: Science 366.6464 (2019), pp. 447–453.
<https://science.sciencemag.org/content/366/6464/447>
- Geiger, G., ***Court Rules Deliveroo Used 'Discriminatory' Algorithm***, Motherboard, January 5 (2021), <https://www.vice.com/en/article/7k9e4e/court-rules-deliveroo-used-discriminatory-algorithm>

Welche Rechenressourcen verbraucht der Algorithmus beim Training?

- Hao, K., ***Training a single AI model can emit as much carbon as five cars in their lifetimes***, in MIT Technology Review, June 6, 2019
<https://www.technologyreview.com/s/613630/training-a-single-ai-model-can-emit-as-much-carbon-as-five-cars-in-their-lifetimes/>

Seien Sie Bias Detektiv

Dieses Experiment dürfen Sie gerne zu Hause ausführen!*

Geben Sie folgende Texte in Google Translate rein, und übersetzen Sie aus dem Englischen, ins Deutsche:

- Englisch: My teacher is clever, he immediately found the solution
- Google Deutsch:
- Englisch: My teacher is clever, she immediately found the solution
- Google Deutsch:
- Englisch: My doctor is clever, she immediately found the solution
- Google Deutsch:

***Hat tip Liad Magen für die Idee zum Experiment.**

FRAGEN?

Danke für Ihre Aufmerksamkeit!

*Rania Wazir, Ph.D.
VDSG/data4good
www.linkedin.com/in/raniawazir*