Themenblock 10 Explainable/Hybrid/Robust AI, 10_01Frage_InterpretableML

Interpretable Machine Learning

Erarbeitet von
Marc Feger M.Sc.

Lernziele 1
Inhalt 2
Einstieg 2
Warum ist Interpretierbare KI wichtig? 2
Interpretierbare KI: Innovation, Ethik und Fortschritt? 2
Zusammenfassung 3
Quellen 3
Weiterführendes Material 4
Disclaimer 4

Lernziele

- Du verstehst die Bedeutung von Interpretierbarkeit in der KI und erkennst ihre entscheidende Rolle beim Aufbau von Vertrauen, bei der Gewährleistung von Verantwortlichkeit und bei der Einhaltung ethischer Standards
- Du kennst die wesentlichen Merkmale interpretierbarer KI-Systeme, wie z. B. Ausdrucksstärke, Transparenz und Übertragbarkeit, und weißt, wie wichtig sie für das Verständnis und die Zugänglichkeit für den Betrachter sind
- Du verfügst über ein Bewusstsein dafür, dass die Interpretierbarkeit in der KI ein wesentlicher Bestandteil ethischer Entscheidungsfindung ist und Innovationen fördert, indem sie den Abbau von Vorurteilen erleichtert, Fairness gewährleistet und zur kontinuierlichen Verbesserung von KI-Anwendungen beiträgt

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

Inhalt

Einstieg

Quelle [1, 3, 5]
Herzlich willkommen!

Heute wollen wir einen Blick in die Welt der künstlichen Intelligenz werfen und uns dabei auf einen wichtigen Aspekt konzentrieren, die Art und Weise, wie wir mit dieser Technologie umgehen und von ihr profitieren: Die Interpretierbarkeit von KI. Im Kern geht es bei KI um Algorithmen, die Entscheidungen treffen, von einfachen bis hin zu hochkomplexen. Da diese Algorithmen jedoch zunehmend verschiedene Aspekte unseres Lebens beeinflussen, besteht ein wachsender Bedarf an Interpretierbarkeit.

Das bedeutet, dass wir nicht nur KI-Systeme brauchen, die Entscheidungen treffen können, sondern Systeme, deren Entscheidungen wir verstehen und rationalisieren können.

Warum ist Interpretierbare KI wichtig?

Quelle [2, 4, 5]
Warum also ist die Interpretierbarkeit in der KI so wichtig? In erster Linie geht es um Vertrauen. In einer Zeit, in der KI-Systeme Entscheidungen treffen, die sich auf vieles auswirken, von persönlichen Empfehlungen bis hin zu bedeutenden geschäftlichen und gesellschaftlichen Ereignissen, ist der Aufbau von Vertrauen unerlässlich. Wenn wir verstehen, wie ein KI-System zu einer Schlussfolgerung kommt, ist es wahrscheinlicher, dass wir den Entscheidungen vertrauen und sie akzeptieren.

Und dann ist da noch der Aspekt der Rechenschaftspflicht. Wenn KI-Systeme kritische Entscheidungen treffen, ist die Möglichkeit, diese Entscheidungen zu interpretieren, der Schlüssel, um die Systeme (und ihre Urheber) zur Verantwortung zu ziehen.

Die Interpretierbarkeit stellt sicher, dass KI-Entscheidungen geprüft, infrage gestellt und gegebenenfalls angefochten werden können.

Interpretierbare KI: Innovation, Ethik und Fortschritt?

Quelle [1, 4]
Außerdem ist die Interpretierbarkeit ein wesentlicher Faktor für Innovation und Fortschritt in der KI. Indem sie verstehen, wie Entscheidungen getroffen werden, können Entwickler und Forscher KI-Algorithmen verbessern und sie effektiver, effizienter und fairer machen.

Insbesondere geht es darum, KI in sensiblen Bereichen, wie der Vergabe von Krediten, Wohnraum oder Arbeitsplätzen, in denen es Vorurteile hinsichtlich verschiedener diskriminierender Merkmale geben kann, zu überprüfen und Missstände messbar zu machen.

Interpretierbare KI ist also ein Kreislauf der kontinuierlichen Verbesserung, der durch Verständnis und Einsicht angetrieben wird.
Betrachten wir nun, was ein KI-System interpretierbar macht. Es geht nicht nur um den Algorithmus selbst, sondern auch darum, wie sein Entscheidungsprozess kommuniziert wird. Ein interpretierbares KI-System liefert Erkenntnisse auf eine Weise, die zugänglich und verständlich ist. Es bricht komplexe, technische Prozesse in klare, verständliche Informationen herunter. Ebenso gibt es eine ethische Dimension zu berücksichtigen. Da sich KI-Systeme zunehmend auf unser Leben auswirken, muss sichergestellt werden, dass diese Systeme nach ethischen Grundsätzen funktionieren.

Die Interpretierbarkeit spielt auch hier eine Schlüsselrolle. Sie ermöglicht es uns, sicherzustellen, dass KI-Entscheidungen ohne Voreingenommenheit und unter Wahrung der Privatsphäre und der allgemeinen Achtung der in der Gesellschaft geltenden Rechte getroffen werden.

Zusammenfassung

Deshalb ist die Interpretierbarkeit von KI nicht nur eine Funktion, sondern eine grundlegende Notwendigkeit für eine sichere Gestaltung von KI als integraler Bestandteil des Alltags. Wenn wir uns die Möglichkeiten der KI zunutze machen, müssen wir dafür sorgen, dass sie nicht nur in ihren Fähigkeiten, sondern auch in ihrer Transparenz und Verständlichkeit wächst.

Vielen Dank für dein Interesse an unserem heutigen Thema zur Interpretierbarkeit von KI.

Quellen

Quelle [1] Yoon, C., Torrance, R., & Scheinerman, N. (2021). Machine learning in medicine: should the pursuit of enhanced interpretability be abandoned?. https://doi.org/10.1136/medethics-2020-107102

Quelle [2] Choudhary, S., Chatterjee, N., & Saha, S. (2022). Interpretation of Black Box NLP Models: A Survey. https://doi.org/10.48550/arXiv.2203.17081

Quelle [3] Tjoa, E., & Guan, C. (2019). A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI https://doi.org/10.1109/TNNLS.2020.3027314

Quelle [4] Abdollahi, B., & Nasraoui, O. (2018). Transparency in Fair Machine Learning: the Case of Explainable Recommender Systems. https://doi.org/10.1007/978-3-319-90403-0_2

Quelle [5] Sokol, K., & Flach, P. (2020). One Explanation Does Not Fit All. https://doi.org/10.1007/s13218-020-00637-y

Weiterführendes Material

Molnar, Christoph. Interpretable Machine Learning: A Guide for Making Black Box Models Explainable., https://christophm.github.io/interpretable-ml-book/

Disclaimer

Transkript zu dem Video „Themenblock 10 Explainable/Hybrid/Robust AI, 10_01Frage_InterpretableML“, Marc Feger.