idw - Informationsdienst Wissenschaft
Nachrichten, Termine, Experten

Pressemitteilung

Technische Universität Chemnitz
Matthias Fejes

30.01.2019
http://idw-online.de/de/news709766

Forschungs- / Wissenstransfer
Informationstechnik
überregional

TECHNISCHE UNIVERSITÄT CHEMNITZ

Machine prejudice

Challenges of accountability and responsibility when dealing with Artificial Intelligence

Threats imposed by Artificial Intelligence (AI) are commonly associated with its supposed superior capabilities. Instead, the most imminent danger arises from overconfidence in the apparent objectivity of machine-based decisions. Wolfgang Einhäuser-Treyer and Alexandra Bendixen, professors at Chemnitz University of Technology (TU Chemnitz), argue for a push towards an explainable and accountable AI

AI is by no means a novel concept. Its theoretical and philosophical roots date back to the early days of computer science in the mid-20th century. Despite its success in domains, such as handwritten digit recognition in the early 1980s, only recently has AI gained the interest of the broader public. Current progress in AI is not primarily related to fundamentally new concepts, but to the availability and concentration of huge masses of data – with „Big Data“ nowadays becoming a field of its own.

The promise: Recognition of hidden patterns and information

AI finds and recognises patterns in huge amounts of data. For example, from many photos of animals that are labelled as either “cat” or “dog”, it can learn to distinguish these species. Importantly, no one needs to teach the AI what specifically differentiates a cat from a dog – the photos and their labels suffice. AI extracts and learns the relevant, discriminative features all by itself. Self-learning AI systems become more interesting when they extract patterns that are not noticed or noticeable to humans. For example, which features could AI discern that may determine creditworthiness among customers?

Consequently, expectations and hopes for potential usage are huge: decisions of high personal and societal relevance shall be transferred to the AI. Examples include credit scores, medical treatments and using video data to predict potentially dangerous behaviour in public areas. The apparently “objective” algorithm is hoped to make “better”, “faster” and – most of all – “fairer” decisions than human judgement, which is prone to prejudice.

An unavoidable weakness: biased training data

This all sounds extremely promising, however: are the AI’s decisions really objective and neutral? This would require the database used for training to be completely unbiased. For example, if a certain subspecies of cat occurs too rarely in the training data, it is likely that the algorithm will not properly recognise those cats. This example is not chosen incidentally: some years ago, a severe AI error of this kind led to a – completely justified – public outcry: the automatic face-recognition system by a large software company classified a group of African Americans erroneously as “gorillas”. How could such a grave error possibly occur? It is highly likely that in the training database, the category “human” was biased towards Caucasians – in other words: the database included far more Caucasian people than people of colour. This sad example demonstrates how biases in databases can lead to dramatic errors in the AI’s judgement, which would

rightfully be considered awful acts of discrimination if performed by a human.

Machine prejudice

If AI errs even for such simple questions (i.e., is a human depicted or not?), what about more complex situations? What about decisions on credit scores, where banks and insurance companies already rely on AI? In this case, the database consists of a collection of information about persons who received credit payments and whether they were able to pay them back. When the AI has "learned" that people who failed to pay back their debts share certain features, it will infer that other people with those features should not be considered creditworthy in the future. As for image classification, the algorithm will necessarily err if the training database was biased. This brings about two issues that are substantially different from the species-classification example: Firstly, no human can detect the error of the AI system, as the "correct" answer is not evident – no one will ever know whether the person who was ruled unworthy of credit by the AI would in fact have paid their credit back. Secondly, in addition to severely affecting the person's dignity, the irreversible credit decision also compromises their opportunity to develop in society – in the worst case, discrimination turns into a threat against the person's existence.

Shifting responsibility and the lack of accountability

Biases in databases that are used to train self-learning algorithms unavoidably lead to prejudice in judging individual cases. The issue of prejudiced judgment is further enhanced if users and executing agents remain unaware of these biases or shift the responsibility to the apparently "objective" AI: "The computer decided based on your data that you are unworthy of a credit – as an employee, there is nothing I can do about this."

One may argue that humans are subject to their own prejudices, preconceptions and biases. While this objection is undeniably valid, humans – as opposed to AI systems – can be confronted with their prejudices and held accountable for their decisions. Human decision makers remain accountable and responsible even if they are not fully aware of their biases and their decision-making processes. For self-learning AI, the biases – the prejudices – are already present in the training database. In addition, for sufficiently complex systems, it is typically impossible to trace the basis of an individual decision – that is, to determine the features that influenced the decision and the rules that connected and weighed these features. This undermines accountability and responsibility for decision-making processes. Accountability and responsibility are, however, the central prerequisites for accepting and complying with others' decisions, thereby forming the foundation on which all just societal and legal systems rest.

Machine biases cannot be corrected

A further rebuttal to the concern about machine biases states that larger and larger amounts of data will eventually render database biases negligible. Yet even the simple example of credit scores shows how biases become irreversible as soon as one blindly follows the AI's decisions. If certain groups do not receive credit, no further data will be collected on whether they would in fact have been worthy of it. The database remains biased, and thus decisions continue to follow prejudiced judgements. This, by the way, is the same mechanism that frequently yields persistent prejudice in humans, which could be readily overcome if they only acquired a larger "database" through more diverse experiences.

The actual opportunity: connecting large sets of data under human control and responsibility

The arguments expressed so far shall by no means imply that using AI is necessarily disadvantageous. On the contrary, there are areas where the detection of connections in large databases bears tremendous advantages for the individual. For example, patients with rare symptoms will greatly benefit if they do not have to rely on the knowledge base of a single physician, but can use the weighed experience of the entire medical field. In an ideal case, AI provides new ideas about how symptoms should be interpreted and how the patient should be treated. Yet even here, before therapeutic decisions for the individual are derived, the risk of database biases needs to be carefully considered – in other words, the

AI generates hypotheses, ideas and suggestions, but the human (physician) remains in control of the eventual decision.

The key challenge: accountable AI

How can the advantages of self-learning AI systems be utilised to the benefit of society and the individual, without suffering from the aforementioned disadvantages? AI researchers and the general public need to raise awareness of the unavoidable biases of databases. Similarly to humans, seemingly objective algorithms also suffer from prejudice, whose sources are frequently hard to trace. Shifting responsibility and accountability for decisions to AI systems is therefore unacceptable. At the same time, research on procedures that uncover the basis of decision-making in AI systems ("explainable AI") needs to be strengthened. Being able to fully trace and comprehend data-processing and decision-making processes is an essential prerequisite for an individual's ability to stay in control of their own data, which is a fundamental right of utmost importance in information-age societies.

The role of schools and universities

Explainable and accountable AI will benefit immensely from the interaction with those academic disciplines that have always dealt with prejudice in decision making – psychology, cognitive sciences and to some extent economics. Schools and universities must embrace the challenge to teach a comprehensive view on AI. In engineering, the sciences, and the computational disciplines, this requires a thorough discussion of the societal consequences of biased data; in the humanities and social sciences, the understanding of technology needs to be strengthened. A solid and broad education in human, societal and technological aspects of self-learning information-processing systems will provide the best qualification to assess the risks and opportunities of AI, to design and construct AI systems and to ensure their responsible use for the benefit of both the individual and the society as a whole.

Keywords: Prof. Dr. Wolfgang Einhäuser-Treyer and Prof. Dr. Alexandra Bendixen

Wolfgang Einhäuser-Treyer (mytuc.org/zndx) studied physics in Heidelberg and Zurich. After receiving his Ph.D. in Neuroinformatics from the Swiss Federal Institute of Technology (ETH) Zurich, postdoctoral stays at Caltech and ETH, including research in the area of machine learning, and an assistant professorship in Neurophysics at the University of Marburg, he joined the faculty of TU Chemnitz in 2015 as a full professor for 'Physics of Cognitive Processes'.

Alexandra Bendixen (mytuc.org/gmrx) studied psychology in Leipzig and Grenoble and obtained her Ph.D. in Cognitive and Brain Sciences from Leipzig University. After a postdoctoral stay at the Hungarian Academy of Sciences in Budapest, her Habilitation at Leipzig University, and an assistant professorship for 'psychophysiology of hearing' at Carl von Ossietzky University of Oldenburg, she joined the faculty of TU Chemnitz as a full professor for 'Structure and Function of Cognitive Systems' in 2015.

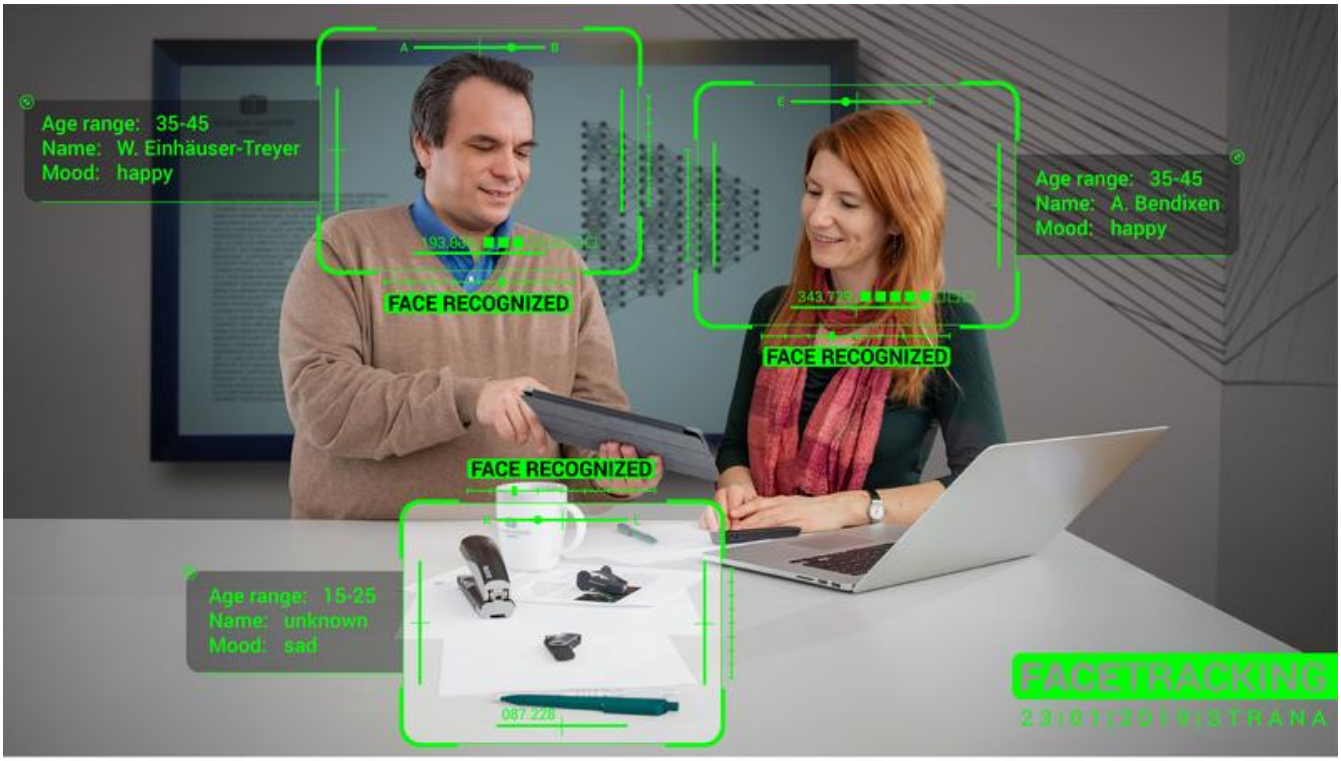
In the 'Sensors and Cognitive Psychology' (mytuc.org/ngwk) study program at TU Chemnitz, they both strive to convey the connection between the technological and human perspective to their students.

(Author: Prof. Dr. Wolfgang Einhäuser-Treyer und Prof. Dr. Alexandra Bendixen / Translation: Jeffrey Karnitz)

wissenschaftliche Ansprechpartner:

Prof. Dr. Wolfgang Einhäuser-Treyer, Physics of Cognitive Processes:
wolfgang.einhaeuser-treyer@physik.tu-chemnitz.de

Prof. Dr. Alexandra Bendixen, Structure and Function of Cognitive Systems:
alexandra.bendixen@physik.tu-chemnitz.de



Prof. Dr. Wolfgang Treuer and Prof. Dr. Alexandra Bendixen make the case for an accountable AI.
Photo: Jacob Müller