[The original, German version of this article appeared on February 27, 2021, as part of the AI column “Aus dem Maschinenraum der KI” in the financial section of the German national daily newspaper Die Welt, p. 16.]

[Translated with www.DeepL.com/Translator (free version) - AI technology made in Europe, see https://en.wikipedia.org/wiki/DeepL_Translator, and subsequently polished and modified by the author.]

Messages from the AI engine room

Nobody is Perfect!

Both, humans and machines have many opportunities to "misclassify" the world around them — they can and should learn from each other.

By Kristian Kersting and Constantin Rothkopf

Artificial intelligence (AI) will change the way people live and work, very much like steam engines and computers did. The current pandemic is accelerating this transformation — from contactless payment to the use of recommender systems on streaming platforms to robots delivering packages.

This transformation, however, also presents us with challenges: For example, a major e-commerce company in the U.S. realized to its regret that an in-house developed AI system used in the HR department had apparently learned to like applications of women less than those of men — a discrimination that neither the company in question nor society can tolerate.

But how do stereotypes and biases enter AI systems? How do inequity and unfairness arise? Do biased people program such AI algorithms on purpose? Do they even intentionally train AI systems with biased data?

The term "bias" plays a central role here. Imagine that ten polling institutes conducted polls on the next election. If all the institutes asked, say, one thousand members of the Christian Democratic Union of Germany (CDU) for their votes, then the predictions about the election outcome would naturally prove to be not representative. This deviation from the actual election result is the so-called bias. It can be very high, even though the ten predictions may appear to be quite consistent with each other and show only a small variance.

Now imagine that each institute instead asks only five randomly selected people, which German party they would vote for if federal elections were held this Sunday? We would certainly get very different answers, i.e., a large variance among the election predictions. Taken together, however, they probably are already very close to the actual election result and, thus, show a small bias.

We encounter bias and variance everywhere. If we feed AI systems with randomly selected texts from the World Wide Web, men are actually mentioned more often than women in connection with the word "science". Unfortunately — and this may come as a surprise — a fundamental law of information processing states that (1) bias and variance can only be reduced to a certain extent, but never to zero, and (2) they are interdependent: Either the

bias is small, but then the variance gets larger, or vice versa. And what is even more astonishing: This law is valid for every information-processing system — also for us humans. The idea that there are people who are always right is a bias.

Recently, Stanford Medical Center deployed an algorithm to determine the order in which to vaccinate employees against Covid-19. The AI system consisted of a few rules and took into account employee-based variables such as age, as well as workplace-based variables and public health guidelines. Due to flaws in the human-designed rules, admins and employees who worked at home ended up being ranked high, while only 7 out of the 1,300 “frontline” medics in residence made it into the list — a debacle reported by the "Washington Post" and the "New York Times."

That we humans make mistakes is a well-known fact. This is almost the definition of the Conditio Humana: "errare humane est” – to err is human. Actually, you can see easily that we often think in clichés and stereotypes. The following puzzle was published by Boston University: " A father and son are in a horrible car crash that kills the dad. The son is rushed to the hospital; just as he’s about to go under the knife, the surgeon says, “I can’t operate — that boy is my son!” — Explain. Not even 25% of the subjects came up with the mom’s-the-surgeon answer.

Even our human, unconscious facial recognition is not unbiased. If we have not had regular social contact with people of other ethnicity by the age of twelve, we will have lifelong difficulties distinguishing their faces. As Osgood Fiedling said in "Some Like it Hot" when it was revealed to him that Daphne was a man (namely, Jack Lemmon): "Well, nobody is perfect!" This is true for humans as well as machines. Consequently, in areas where we want to avoid bias, we need to program and train AI systems in a way that they do not produce undesirable results.

But be careful! Human judgments may be driven by similar algorithms, and often we are not even aware of this. For example, research shows that judges rule more generously after a break or a meal. That's not a bias we want as a society.

What can we do? Expose biases and work hard to objectify and reduce them. AI systems are important for that. They also hold up a mirror to us and help us uncover and, hopefully, reduce our own biases. This is an important step in eliminating bias in society.

Kristian Kersting is Professor of AI and Machine Learning at TU Darmstadt, co-director of the Hessian Center for AI (hessian.ai), and winner of the "German AI Award 2019". His AI column "Aus dem Maschinenraum der KI" appears monthly in the German national daily newspaper “Die Welt.” **Constantin Rothkopf** is a Professor of Cognitive Science at TU Darmstadt, director of its Centre for Cognitive Science, and a member of hessian.ai. Both have co-authored together the book "Wie Maschinen Lernen."