Programming **fairness**

In the future, it will be more and more common for computers to make decisions about human beings – whether they are granting loans or assessing applicants. However, it happens occasionally that the automated systems that are already in use discriminate against certain groups of people. **Niki Kilbertus** and **Bernhard Schölkopf**, researchers at the **Max Planck Institute for Intelligent Systems** in Tuebingen, want to change this by developing fair algorithms.

TEXT **TIM SCHRÖDER**

Several blackboards are dotted around the hallways of the Max Planck Institute for Intelligent Systems in Tuebingen. Scientists walking past can use these to note down their thoughts. They are also places for researchers who happen to meet in the corridors to discuss new ideas. "It's really helpful," says Niki Kilbertus, "because I have to develop so many ideas right now." He is referring to concepts that go far beyond those he has dealt with in the past. Niki Kilbertus has degrees in mathematics and physics. He is familiar with the rigorous formal methods for solving complex matters, and he knows how to program algorithms. For his doctorate, however, he has ventured beyond the boundaries of formal languages. He is addressing a question that has been the subject of heated public debate for a while now: whether and to what extent algorithms can be fair.

Balanced calculation: algorithms must not discriminate against particular groups of people, such as women or men.

Photo: Alamy

Computers are sober machines that are incorruptible and are never wrong – or so one might think. Nevertheless, debate about computer algorithms discriminating against people flared up at the end of 2018. It had become known that a large online retailer was planning to use a computer for pre-screening applicants, and that it had emerged even in the trial phase that the computer was more likely to reject applications received from women than those from men. This led to an outcry in the media, not least because experts envisage that in the future, computers will be making decisions about human beings more and more often, using vast amounts of data that are now available. It would be scandalous if these computers favored or discriminated against particular groups of people.

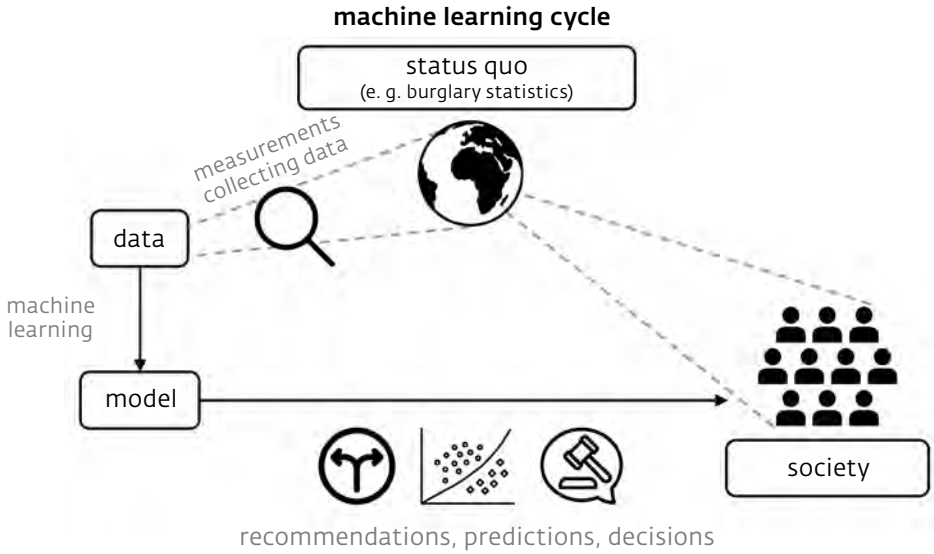
Initiatives have been formed around the world in light of such scenarios, promoting fairness in artificial intelligence. This is not their only concern, they also demand that companies must be held accountable for their algorithms, calling for responsibility and accountability. Critics also expect transparency when it comes to how and why particular decisions are made based on calculation specifications. The community is called FAT: fairness, accountability, transparency.

ANY DISCRIMINATION MUST BE ANALYZED CAREFULLY

Niki Kilbertus is a doctoral student in the Cambridge-Tuebingen-Program run by the Max Planck Society, and works at the Max Planck Institute in Tuebingen as well as at Pembroke College in Cambridge. He and his colleagues are exploring the part of FAT that can be translated into technology: the researchers strive to teach fairness to algorithms. To do so, they rely on machine learning. This means that they enable the computer to gradually improve through learning and experience.

It is a challenging task. So far there is no technical solution that allows for fairness to be instantly realized across all applications. It is likely that this is not going to change in the future either: "There can be no solution that answers every possible question. Each situation is different, and needs to be analyzed separately," says Kilbertus. "And the second step is at least as elaborate: we need to find a mathematical description for the real-world problem."

The algorithmic analysis process always follows a clear formula: data is collected and fed into an analysis program, whose algorithms will then issue a recommendation such as "applicant not suitable". To ensure that this automated process is always fair in the future, it would need to be improved on both the data and the output side. >



"First, you need to examine which data is collected, and how, and then check which decisions are finally issued by the computer, and the reasons behind them," explains Niki Kilbertus. These analyses always touch on data protection issues as well. "This is always about sensitive personal data, after all. This means that we also need to find solutions for analyzing data without disclosing it."

The young researcher is part of the working group led by Bernhard Schölkopf, Director of the Max Planck Institute for Intelligent Systems in Tuebingen. Both researchers deal with the basic issue of causality in machine learning: the question as to what extent computers can draw meaningful conclusions between different aspects. This very question is also a central aspect when it comes to the issue of fairness.

Niki Kilbertus has an example: the granting of loans and checking of creditworthiness. Things can get critical if simple algorithms are used that are based purely on correlations "if X, then Y relationships". If residents of a particular city district were less likely in the past to pay back their loans, the algorithm could, for example, use the place of residence as an indicator for future applicants. The algorithm does not understand in this context that the place of residence is unlikely to have a direct causal impact on creditworthiness, and that there are likely to be other, more relevant factors at play.

THE ALGORITHM WOULD NEED TO VERIFY ITS ASSUMPTIONS

Systems are often trained using historic data about credit repayment behavior. Personal data, such as the place of residence of a new applicant, are then used to calculate their creditworthiness: how likely is it that the applicant is going to pay back their loan? However, the computer may unintentionally discriminate against individuals, since a creditworthy person might live in an area to which the algorithm attributes a poor reputation.

"The algorithm would basically need to be equipped to verify its as-

Above Real-life feedback: a machine learning algorithm solves a question, such as which neighborhoods are likely to experience a lot of burglaries. For this purpose it is trained using data about the status quo, for example, data regarding the economic and social situation, and crime rates. The algorithm then develops a model, based on which it issues predictions, decisions, or recommendations. If criminal acts are then prevented in a particular neighborhood due to increased police presence, society and thus the state of the world is changed. The algorithm then has to adjust the model based on new data.

Right Freedom for ideas: Mateo Rojas-Carulla, Niki Kilbertus, and Nadine Rüegg (left to right) are discussing various aspects of artificial intelligence.

Graphic: MPI for Intelligent Systems

Photo: P. Junker / MPI for Intelligent Systems

sumptions on a regular basis," says Niki Kilbertus. For example, by using specific criteria to occasionally grant a loan to an individual that is initially not rated as creditworthy. Economists refer to this approach as *explore versus exploit*. To *explore* is to test new solutions in this context. To *exploit*, on the other hand, is to utilize an existing approach as effectively as possible, to avoid the effort required for a new development, for example. In the case of granting loans, the *explore* approach would provide for loans to be occasionally approved contrary to the original rules. If the person turns out to pay back the loan after all, the algorithm has to be adjusted and improved.

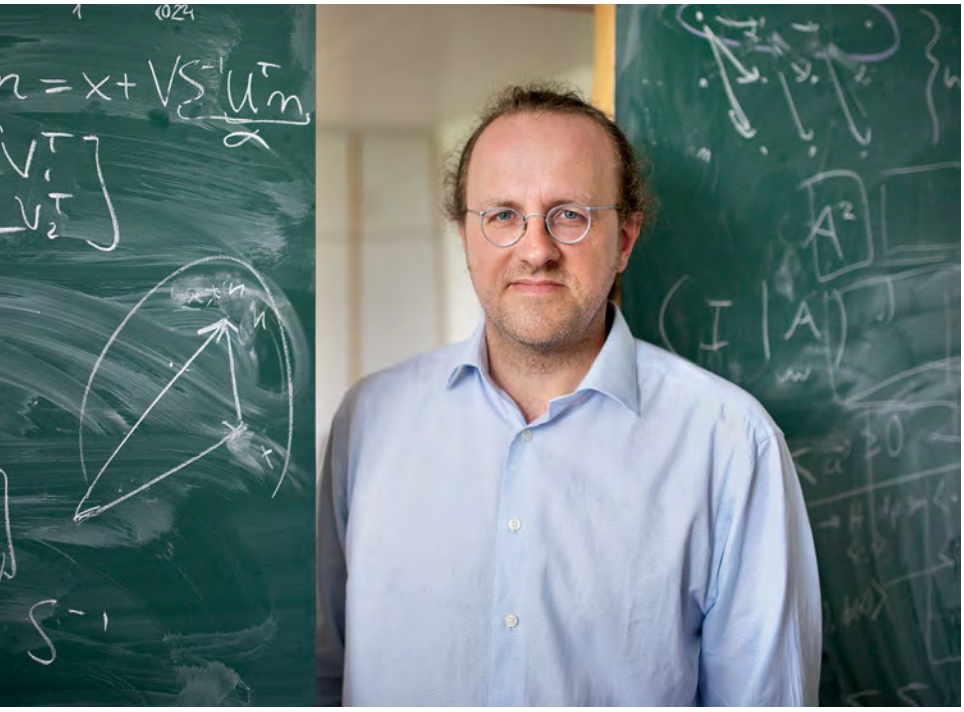
This example shows that in some cases, the data that is currently used is not sufficient in order to make fair decisions. Niki Kilbertus: "When it comes to fairness, the challenge is to discover and understand the true causal structure underlying the data as far as possible." This approach is intended to prevent an algorithm from considering data as being merely a table of figures among which it can happily establish correlations that are often likely to have nothing to do with the true causal connections.

TRUE CAUSAL CONNECTIONS ARE KEY

Niki Kilbertus uses a far-fetched example to illustrate this point. An algorithm could detect that people who are more likely to repay their loans tend to be tidier and more likely to stick felt gliders under the chairs in their home. A calculation specification does not per se understand correlations, so it may conclude that it is a good idea to give a pack of felt gliders to new borrowers in order to increase their creditworthiness. "We need to understand the causal connections to be able to ask any meaningful questions, such as whether the individual would have repaid the loan if they had lived elsewhere," says Niki Kilbertus. By contrast, incorrect causalities such as the link between possessing felt gliders and repayment behavior could be eliminated.

Simpson's paradox is another illustration of the importance of correct identification of causal connections. It was named after the British statistician Edward H. Simpson, who demonstrated in the 1950s that a particular combination of data from different groups, such as women and men, will give rise to seemingly paradox situations. One classic example of a Simpson's paradox dates back to 1973. There was a sense of outrage at the time, when out of the women who had applied for a place at the University of California in Berkeley in the U.S., a smaller percentage had been accepted to take up studies than was the case for male applicants.

It turned out, however, that this was not a case of discrimination. An analysis of the data conducted later on showed that more women had registered for extremely popular subjects in the fields of humanities and social sciences. They were therefore on the whole rejected more often than men, who more frequently applied for less sought-after subjects such as chemistry or engineering. The paradox was that



Teaching machines: Bernhard Schölkopf teaches algorithms to recognize true causal connections among data. This also promotes fairness of decisions made by computers.

for each Department in isolation the percentage of women accepted for most courses of study was even higher than that of accepted male applicants. However, across all disciplines, the proportion of male applicants that were accepted was slightly higher. "To understand the causal connection correctly, we first need to have the right data and to know which Department the women applied for. Gender has an impact on the choice of subject. Only then are we able to correctly interpret the situation," says Kilbertus.

Bernhard Schölkopf even goes a step farther when it comes to the issue of correct causalities. He also addresses the question "where does fairness begin?" Does it begin with asking whether African-American applicants have the same opportunities in the U.S. as their white competitors? "Or do you have to go back further, and take into account the fact that colored children do not have the same educational opportunities as white children, and that this has an impact on their entire résumé and their future job prospects?" Schölkopf raises the question of whether aspects like this should also be included to create truly fair calculation specifications. An algorithm with multiple levels of fairness so to speak.

EVEN PEOPLE ARE NOT ALWAYS FAIR

Nevertheless, he also points out that we should not get carried away too easily. "People make decisions about other people every day. It is often completely unclear why an individual makes a certain decision. At the same time, we demand that an algorithm must always make decisions that are one hundred percent correct and fair." At the end of the day, studying fairness in machines also leads to the insight that human beings are also not always fair. Quite simply, we make mistakes, for example due to a lack of information or experience. Much like machines. A dermatologist for example, who spent their entire career screening only light-skinned people for skin cancer, may be more likely to mis-diagnose dark-skinned people, and to possibly overlook a tumor. This means that dark-skinned people would not receive the same quality of treatment as light-skinned people. However, the supposed unfairness can be explained by the fact that the doctor is lacking experience in working with particular groups of patients.

Whether an algorithm is going to make fair decisions in the future also depends on how companies or people in general define or perceive fairness. "The ultimate goal is that an algorithm should always make the right decision," says Niki Kilbertus. "But first we have to clarify for each case what right even means."

There is something else that needs to be taken into account: decisions made by computers can have an active impact on the world. One example of this are modern programs that use data about cases of burglary and theft in a city to determine which neighborhoods are most likely to experience further burglaries. The algorithm's statement will then lead to increased police presence in these areas, and possibly to an increase of crimes detected. It is conceivable, however, that it was not actually the case that more criminal offenses were committed in these neighborhoods. Possibly, it was simply that more crimes were detected because the police patroled these areas more frequently.

"The decisions based on an algorithm can therefore lead to wrong conclusions. We refer to this as a feedback loop, in which the algorithm has an impact on real life." Another similar example are traffic warnings issued by navigation services. If a road is busy, the service will recommend switching to other routes, and after a short while,

Photo: Wolfram Scheible

traffic will build up on these routes. This is why Niki Kilbertus points out that it is important to keep not only fairness in mind, but also possible feedback effects.

In the context of his doctorate, he has analyzed a number of cases of discrimination and made initial attempts to describe the respective problems mathematically. "Programming takes at least as much time again as analyzing each case." And yet for the time being, he and his colleagues are focusing mostly on analysis. "Each algorithm works based on specific criteria that are used to derive a statement. We are now trying to find out where the criteria fail; to find the weak point."

An interesting question in this context is why the algorithm by the online retailer mentioned above seems to have discriminated against female applicants. It is unknown how the algorithm works in any detail, but Kilbertus has a hunch. "It is fair to assume that they set out to analyze the applications in an anonymized form." So the system would not have known the applicants' names or their gender. Nevertheless, the system failed. "Other studies have shown that women tend to mention social commitment or related activities in their applications, while men are more likely to signal dominance and to come across as more competitive," Kilbertus explains. "These aspects could be the very characteristics that the company is looking for." As Bernhard Schölkopf points out: "It is incorrect to assume that we can generally achieve fairness, by withholding certain details such as gender from the algorithm." This *fairness by unawareness* approach is far from infallible.

Whichever way fairness might be taught to algorithms, it is a rather soft criterion compared to the purely mathematic formalisms Niki Kilbertus used to deal with in the past. While he does appreciate being able to produce clear evidence for or against a statement, he also finds this new aspect very rewarding. "It is interesting to work on an issue of such social relevance," he says. He soon realized that knowledge in the fields of mathematics and computer sciences is not enough for his research work. So he decided to learn about social sciences and legal issues. He hopes that he is now well-equipped for finding truly fair algorithms. ◂

SUMMARY

- **Algorithms make decisions about human beings more and more often. Discrimination occurs time and again in this context, not least because incorrect causal connections are established from data, such as links between an individual's place of residence and their creditworthiness.**
- **In their endeavor to teach fairness to algorithms, the researchers at the Max Planck Institute for Intelligent Systems in Tuebingen analyze the data used by the calculation specification and how decisions are made in each individual case.**
- **Based on their findings concerning true causal connections that provide answers to questions such as "Which person is going to pay back a loan?", the researchers are drafting mathematical descriptions for the respective issues at hand.**