BOOK REVIEW

The Ethical Algorithm by Michael Kearns and Aaron Roth (2020) 256pp. £19.00 (hardback) Oxford University Press, New York, ISBN: 13 9780190948207

Some 20 years ago, the governor of Massachusetts agreed to make data summarizing hospital visits for every state employee available to academic researchers. To protect people's privacy, the governor promised that the records would be anonymized before being released. All explicit patient identifiers were removed, including names, addresses and social security numbers. However, Latanya Sweeney, a PhD student at MIT, did not feel convinced by the governor's claim that this would be sufficient to guarantee anonymity. She spent $20 on a copy of the voter rolls for the city of Cambridge, which contained the name, birth date, address and sex of every Cambridge voter, include the governor's. It turned out that no more than six patients had the same birthdate as the governor, and only one was a male living in the governor's zip code. It was a piece of cake for Ms Sweeny to find the governor's medical records in the 'anonymized' data set. To make her point, she sent a copy of the governor's records to his office.

In *The Ethical Algorithm: The Science of Socially Aware Algorithm Design,* Michael Kearns and Aaron Roth discuss this and numerous other anecdotes they believe illustrate morally problematic uses of algorithms. For instance, did you know that the list of movies you rate on Netflix can be used to uniquely identify you about 99% of the time? By combing data from different sources, it is almost always possible to figure out who says, thinks or prefers what online. This may, of course, not be what we want, or what we ought to strive for, so a discussion of ethical aspects of algorithms is no doubt welcome.

The Ethical Algorithm focuses on two moral values: privacy and fairness. The reason for this is that privacy and fairness 'are perhaps the two areas of research on ethical algorithms that have received the most scientific attention and have the most mature literatures, theories, and experimental methodologies' (p.169). The point of departure for the discussion of privacy is straightforward: in many cases the best way to address whatever concerns people may have about privacy violations is to make sure they never arise. By using numerous helpful examples, Kearns and Roth point out that in many cases data cannot be anonymized by simply removing a few pieces of information here and there. Clever hackers and computer scientists will often be able to restore the missing information by using various tricks, some of which are discussed in the book. So, instead of weighing the pros and cons of releasing information that may not be fully anonymous, Kearns and Roth argue that we should design algorithms in a manner that guarantees that it is impossible (in a theoretical sense) to restore information meant to be kept confidential. It turns out that one of the best algorithms for doing this is based on a method called 'differential privacy'.

Kearns and Roth are theoretical computer scientists, so they claim to be doing science. To be more precise, they write that their aim is to give an accessible overview of what they consider to be the science of ethical algorithms (p.21). They claim that this is a science in its infancy, a science that is likely to develop rapidly in coming years. However, the authors' discussion of privacy suggests that what they are doing should not be presented as a science in its infancy, but as a rather mature and well-established science. Kearns and Roth do not reflect much on what privacy is, why privacy might be important or under what conditions (if any) violations of privacy might be justified. Their goal is simply to discuss how computer scientists can write algorithms that make it difficult for unauthorized people to access sensitive information. This is hardly a novel topic. It is, of course, possible that some of the algorithms are somewhat novel from a technical point of view, but from broader perspective, this is just a general audience book on computer science.

I agree with Kearns and Roth that privacy is an important ethical value, and I also agree that it is often desirable to use algorithms that guarantee (in a strict mathematical sense) that hackers and computer scientists cannot restore deleted information in the manner Latanya Sweeney did. However, I am somewhat sceptical that developing such algorithms should be a central task for a 'science of ethical algorithms'. As I see it, privacy is one of several ethical values, which often has to be balanced against other, conflicting values. If we, for instance, can solve a serious crime by intruding on someone's privacy, that might be a good thing. Privacy is not an absolute value that should be respected under all circumstances, come what may. To make things worse, reasonable people can disagree on how much privacy it would be worth sacrificing in order to achieve something else we care about (crime prevention). Kearns and Roth have little to say about how such conflicts between competing ethical values should be handled. Their view seems to be that we have to wait until 'society' has been able to settle this and other similar disputes before we can develop algorithms that help us make more ethical decisions. This might, of course, be true, but for anyone familiar with the literature on the ethics of technology, this should be a reason for doubting the value of 'the science of ethical algorithms'. If all ethical values have to be defined externally by the programmer, then 'the science of ethical algorithms' can help us solve only problems that are fairly trivial from an ethical point of view.

Although I have sympathy for the overall goal of Kearns and Roth's project, I worry that they perhaps overstate the novelty of their work. Another worry is that the authors do not seem to be familiar with basic moral theories, or any of the ideas discussed in the vast literature on computer and information ethics. One may compare this to a book on economics written by authors with high school knowledge of economics, or a book on psychology written by authors with a common-sense understanding of psychology. I am not saying that such books must be bad. My point is merely that the absence of formal expertise in an area is a red flag that could be a reason for thinking twice before one decides to believe a claim. (Full disclosure: I am not a computer scientist. I took a couple classes in computer science in college, but my formal training is in philosophy. So, any claims I make about computer science are based on limited knowledge of the subject.)

The author's discussion of algorithmic fairness offers additional reasons to remain sceptical of the novelty of the material presented in the book. Kearns and Roth's main example will be familiar to many readers: in 2018, Amazon developed a machine learning algorithm for evaluating résumés submitted by applicants applying for software engineering jobs. Unfortunately, the project had to be abandoned when Amazon discovered that the algorithm penalized résumés containing words suggesting the applicant might be a woman, such as women's soccer team or the name of a prominent all-women college. It turned out that the source of the problem was not the algorithm itself, but the data set used for training the algorithm. The input used for training the system was not entirely unbiased and neutral, so the machine learning algorithm quickly learned to pick up on some of the implicit biases preset in the initial data set.

In the Amazon example, we all agree that the machine learning algorithm led to a morally undesirable outcome. No one believes that job applicants should be rejected because they are women. So, if this is what the algorithm is doing, we should either revise the algorithm or evaluate the applications manually. However, as Kearns and Roth are aware, it is sometimes very difficult to specify how a fair algorithm should behave. There is no uncontroversial, mathematically precise definition of fairness that we can teach an algorithm to mimic. Kearns and Roth elaborate on this at length in a detailed (but not very informative) discussion of algorithms designed to evaluate loan applications. A possible and very simple definition of fairness could be statistical parity: a loan application should be equally likely to be approved regardless of the applicant's gender or race, etc. The problem is that this does not take into account that some subgroups are, on average, more likely to repay their loans. So, another definition of fairness could be that a fair algorithm should take subgroup specific variations into account. However, in doing so, the algorithm may end up approving a higher proportion of applications submitted by, for instance, women. Would this alternative algorithm be more or less fair than the one that seeks to achieve statistical parity?

Kearns and Roth note that no one seems to know the answer to this and other basic questions about fairness. However, what surprises me is that they nevertheless remain optimistic about the possibility of providing a useful, mathematically precise, and uncontroversial definition of concepts as complex as fairness. They optimistically claim that the discussion of fairness is about 15 years behind that of privacy. If we just wait a couple of years, we will find out if it would be unfair to reject the same proportion of loan applications submitted by different subgroups. My view is that this is naïve. Readers familiar with what moral philosophers have written on fairness over the centuries may be inclined to share my pessimism. Aristotle insists on a formal definition of fairness according to which we should 'treat like cases alike',[1] but he has little to say about what would make two cases similar, or about how we should treat cases that are not fully similar. Others have tried to account for fairness in terms of giving people equal opportunities, or by focusing on desert and merit. What makes Kearns and Roth believe that these ancient debates are likely to be resolved once and for all in the near future? Their optimism arguably has little warrant.

This brings me back to another general problem with the book, namely that Kearns and Roth seem to be largely unfamiliar with the academic literature on ethics. Kearns and Roth note that many ethical concepts have no precise definitions. This includes many of the other ethical concepts they consider to be central, such as transparency, accountability and morality (p.170). However, this does not stop the authors from making bold claims about a bright future in which moral machines are able to make more ethical decisions once these conceptual difficulties have been sorted out. But perhaps the truth is that the division of labour between computer scientists and philosophers envisioned by Kearns and Roth is untenable. It may be naïve to think that it is possible to define important ethical values in a manner that makes them mathematically tractable, without taking a stance in century-old philosophical debates on fundamental human values.

Despite the concerns outlined here, there are many things to like about *The Ethical Algorithm*. It is elegantly written, easy to read and full of entertaining examples. I believe this book will appeal to readers looking for an accessible overview of some of the ethical issues that may arise when computer scientists try to build machines that help people make more ethical decisions or make decisions on behalf of humans. Anyone can read this book, no background knowledge is required. My concerns with the book are mostly related to the authors' unwillingness to acknowledge that their project may not be as novel as they think, and that many ethical issues faced by computer scientists are far deeper and harder than Kearns and Roth acknowledge. Ethics is not easy and common sense will not suffice for solving problems with which people have struggled for thousands of years.

Martin Peterson
Department of Philosophy, College of Liberal Arts
Texas A&M University
College Station TX
martinpeterson@tamu.edu

[1] Aristotle, *Nicomachean Ethics,* volume 3, 1131a10–b15.