Quanta magazine

How to Force Our Machines to Play Fair

The computer scientist Cynthia Dwork takes abstract concepts like privacy and fairness and adapts them into machine code for the algorithmic age.

By Kevin Hartnett

Jessica Kourkounis for Quanta Magazine

Theoretical computer science can be as remote and abstract as pure mathematics, but new research often begins in response to concrete, real-world problems. Such is the case with the work of Cynthia Dwork.

Over the course of a distinguished career, Dwork has crafted rigorous solutions to dilemmas that crop up at the messy interface between computing power and human activity. She is most famous for her invention in the early to mid-2000s of "differential privacy," a set of techniques that safeguard the privacy of individuals in a large database. Differential privacy ensures, for example, that a person can contribute their genetic information to a medical database without fear that

anyone analyzing the database will be able to figure out which genetic information is hers — or even whether she has participated in the database at all. And it achieves this security guarantee in a way that allows researchers to use the database to make new discoveries.

Dwork's latest work has a similar flavor to it. In 2011 she became interested in the question of fairness in algorithm design. As she observes, algorithms increasingly control the kinds of experiences we have: They determine the advertisements we see online, the loans we qualify for, the colleges that students get into. Given this influence, it's important that algorithms classify people in ways that are consistent with commonsense notions of fairness. We wouldn't think it's ethical for a bank to offer one set of lending terms to minority applicants and another to white applicants. But as recent work has shown — most notably in the book "Weapons of Math Destruction," by the mathematician Cathy O'Neil — discrimination that we reject in normal life can creep into algorithms.

Privacy and ethics are two questions with their roots in philosophy. These days, they require a solution in computer science. Over the past five years, Dwork, who is currently at Microsoft Research but will be joining the faculty at Harvard University in January, has been working to create a new field of research on algorithmic fairness. Earlier this month she helped organize a workshop at Harvard that brought together computer scientists, law professors and philosophers.

Quanta Magazine spoke with Dwork about algorithmic fairness, her interest in working on problems with big social implications, and how a childhood experience with music shaped the way she thinks about algorithm design today. An edited and condensed version of the interview follows.

QUANTA MAGAZINE: When did it become obvious to you that computer science was where you wanted to spend your time thinking?

CYNTHIA DWORK: I always enjoyed all of my subjects, including science and math. I also really loved English and foreign languages and, well, just about everything. I think that I applied to the engineering school at Princeton a little on a lark. My recollection is that my mother said, you know, this might be a nice combination of interests for you, and I thought, she's right.

It was a little bit of a lark, but on the other hand it seemed as good a place to start as any. It was only in my junior year of college when I first encountered automata theory that I realized that I might be headed not for a programming job in industry but instead toward a Ph.D. There was a definite exposure I had to certain material that I thought was beautiful. I just really enjoyed the theory.

You're best known for your work on differential privacy. What drew you to your present work on "fairness" in algorithms?

I wanted to find another problem. I just wanted something else to think about, for variety. And I had enjoyed the sort of social mission of the privacy work — the idea that we were addressing or attempting to address a very real problem. So I wanted to find a new problem and I wanted one that would have some social implications.

So why fairness?

I could see that it was going to be a major concern in real life.

How so?

I think it was pretty clear that algorithms were going to be used in a way that could affect individuals' options in life. We knew they were being used to determine what kind of advertisements

to show people. We may not be used to thinking of ads as great determiners of our options in life. But what people get exposed to has an impact on them. I also expected that algorithms would be used for at least some kind of screening in college admissions, as well as in determining who would be given loans.

I didn't foresee the extent to which they'd be used to screen candidates for jobs and other important roles. So these things — what kinds of credit options are available to you, what sort of job you might get, what sort of schools you might get into, what things are shown to you in your everyday life as you wander around on the internet — these aren't trivial concerns.

Your 2012 paper that launched this line of your research hinges on the concept of "awareness." Why is this important?

One of the examples in the paper is: Suppose you had a minority group in which the smart students were steered toward math and science, and a dominant group in which the smart students were steered toward finance. Now if someone wanted to write a quick-and-dirty classifier to find smart students, maybe they should just look for students who study finance because, after all, the majority is much bigger than the minority, and so the classifier will be pretty accurate overall. The problem is that not only is this unfair to the minority, but it also has reduced utility compared to a classifier that understands that if you're a member of the minority and you study math, you should be viewed as similar to a member of the majority who studies finance. That gave rise to the title of the paper, "Fairness Through Awareness," meaning cross-cultural awareness.

In that same paper you also draw a distinction between treating individuals fairly and treating groups fairly. You conclude that sometimes it's not enough just to treat individuals fairly — there's also a need to be aware of group differences and to make sure groups of people with similar characteristics are treated fairly.

What we do in the paper is, we start with individual fairness and we discuss what the connection is between individual fairness and group fairness, and we mathematically investigate the question of when individual fairness ensures group fairness and what you can do to ensure group fairness if individual fairness doesn't do the trick.

What's a situation where individual fairness wouldn't be enough to ensure group fairness?

If you have two groups that have very different characteristics. Let's suppose for example that you are looking at college admissions and you're thinking about using test scores as your admission criterion. If you have two groups that have very different performance on standardized tests, then you won't get group fairness if you have one threshold for the standardized-test score.

This is related to the idea of "fair affirmative action" you put forward?

In this particular case, our approach would boil down, in some sense, to what's done in several states, like Texas, where the top students from each high school are guaranteed admission to any state university, including the flagship in Austin. By taking the top students from each different school, even though the schools are segregated, you're getting the top performers from each group.

Something very similar goes into our approach to fair affirmative action. There's an expert on distributive justice at Yale, John Roemer, and one of the proposals he has made is to stratify students according to the educational level of the mother and then in each stratum sort the students according to how many hours they spend each week on homework and to take the top students from each stratum.

Why wouldn't it work to sort the entire population of students by the amount of time they spend on their homework?

Roemer made a really interesting observation that I found very moving, and that is: If you have a student from a very low-education background, they may not even realize it's possible to spend a large number of hours studying per week. It's never been modeled for them, it's never been observed, nobody does it. It may not have even occurred to the student. That really strikes a chord with me.

What is it that you find so moving about that?

I had an interesting experience in high school. I'd started playing the piano at the age of about six, and I dutifully did my half-hour of practice a day. I was fine. But one time — I guess freshman year of high school — I passed by the auditorium and I heard somebody playing a Beethoven sonata. He was a sophomore, and I realized that you didn't have to be on the concert-giving scale to play much, much better than I was playing. I actually started practicing about four hours a day after that. But it had not occurred to me that anything like this was possible until I saw that someone who was just another student could do it. I think probably this is why Roemer's writing struck such a chord with me. I'd had this experience in my own very enriched life.

Your father, Bernard Dwork, was a mathematician and a longtime faculty member at Princeton, so in a sense you had an example to follow — as a scholar if not as a piano player. Did his work inspire yours in any way?

I don't remember his work directly inspiring my interest in computer science. I think growing up in an academic household as opposed to a nonacademic household gave me a model for being deeply interested in my work and thinking about it all the time. Undoubtedly I absorbed some norms of behavior so that it seemed natural to exchange ideas with people and go to meetings and listen to lectures and read, but I don't think it was mathematics per se.

Did that lesson about practice and the piano influence your approach to your research? Or, to put it another way, did you have experiences that taught you what it would take to be successful in computer science?

When I finished my course requirements in graduate school and I started to wonder how I could do research, it turned out that a very famous computer scientist, Jack Edmonds, was visiting the computer science department. I asked him, "How did your greatest results happen? Did they just come to you?" He looked at me, and stared at me, and yelled, "By the sweat of my brow!"

Is that how your best results have come to you?

It's the only way.

You've said that "metrics" for guiding how an algorithm should treat different people are some of the most important things computer scientists need to develop. Could you explain what you mean by a metric and why it's so crucial to ensuring fairness?

I think requiring that similar people be treated similarly is essential to my notion of fairness. It's clearly not the entire story surrounding fairness — there are obviously cases in which people with differences have to be treated differently, and in general it's much more complex. Nonetheless, there are clearly also cases in which people who should be viewed as similar ought to be treated similarly. What a metric means is that you have a way of stating a requirement about how similarly two different people — any two different people — can be treated, which is accomplished by limiting

the amount by which their treatment can differ.

You mentioned previously that you consider this work on fairness a lot harder than your work on privacy, in large part because it's so hard to come up with these metrics. What makes this so hard?

Imagine presenting the applications of two students to a college admissions officer. These students may be quite different from one another. Yet the degree to which they'd be desirable members of the student body could be quite similar. Somehow this similarity metric has to enable you to compare apples to oranges and come up with a meaningful response.

How does this challenge compare to your earlier work on differential privacy?

I think this is a much harder problem. If there were a magical way of finding the right metric — the right way of measuring differences between people — I'd think we had gotten somewhere. But I don't think humans can agree on who should be treated similarly to whom. I certainly have no idea how to use machine learning and other statistical methods to get a good answer to it. I don't see how to avoid dealing with the fact that you need different notions of similarity, even for the same people, but for different things. For example, discriminating in advertising for hair products makes perfect sense in a way that discriminating in advertising for financial products is completely illegal.

When you frame it like that, it seems like a monumental task. Maybe even impossible.

I view this as a "sunshine" situation; that is, the metric that's being used should be made public and people should have the right to argue about it and influence how it evolves. I don't think anything will be right initially. I think we can only do our best and — this is the point that the paper makes very strongly — advocate sunshine for the metric.

This article was reprinted on Wired.com.