Social policy in the age of algorithms: An interview with Jon Kleinberg

Institute advisor Jon Kleinberg on engaging in dialogue with our technology to confront hidden bias and make smarter choices

***For All* Spring 2022** | March 29, 2022

AUTHOR

Jeff Horwich

Senior Economics Writer

You might think Jon Kleinberg, as one of the world's leading authorities on computer algorithms, would only surf the web in "incognito" mode.

"I have colleagues who turn all that stuff off, and that's a very reasonable decision," said Kleinberg. "But I personally found it was simply hard to navigate the world online if I had everything switched off." And of course, each click is an opportunity for a prolific researcher. "I tend to try to figure out why I'm seeing what I'm seeing. What am I going to see in the future as a result of taking a given step?"

Use the technology; learn from the experience; adapt and repeat. Kleinberg brings the same philosophy to his wide-ranging professional work—and a certain amount of faith: With all he knows about the risks and promise of advanced computing, he believes we can ultimately employ it as a force for good.

Kleinberg is the Tisch university professor of computer science at Cornell University. Some of his earliest research in the late '90s laid the conceptual groundwork for the now-dominant Google search engine. He helped establish the modern study of networks—the science of interconnectedness and spread, whether of ideas, illnesses, or financial panics. Along the way, his career has traced the rise of the internet, social media, and the unseen strings of ones and zeros that now permeate many aspects of life.

Kleinberg is in familiar territory as an advisor to the Opportunity & Inclusive Growth Institute, having partnered often with economists—as well as sociologists, doctors, and legal scholars. He dove deep with *For All* into a potent research focus: the potential for modern computing to reinforce our biases, but also to reveal them. This can help us make smarter economic and social policy—if we are willing to truly listen to what our computers are telling us.

Algorithms: A gateway to our hidden biases

Thanks in large part to Facebook and the last two U.S. presidential elections, the word "algorithm" has become a household term. What do you like to give people as a general, working definition?

I think of an algorithm as any procedure that's structured and that can be followed to solve a problem. Your GPS, when it wants to find the shortest drive to your destination, uses an algorithm to do that. Addition is an algorithm. Long division is an algorithm. There are a lot of analogies between algorithms and recipes that we use in cooking.

We've had algorithms for much longer than we've had computers. I think that's important because these terms have a way of isolating the concept, making it seem somehow weird and distinct from the rest of our lives, but it's really blended through our lives. Any time a person, an organization, or a machine carries out a structure or procedure to solve a problem, they're running an algorithm.

You've done a lot of thinking about how algorithms—which have no soul or opinions of their own, as far as I'm aware—can be biased. What's an example of how that happens?

The simplest way would be that you have a procedure that was made up by people who had a bias that they were trying to act on. And now, all the algorithm is doing is formalizing that bias procedure.

Heather Ainsworth for Minneapolis Fed

But I think the more subtle way this happens is with a large, emerging category of algorithms that have become quite powerful over the past 20 to 30 years, called "machine learning algorithms." The idea with machine learning algorithms is that there are a lot of problems that we want to solve that we don't actually know how to write down the rules for. We, as humans, can solve them. But we don't really know *how* we solve them.

The problem comes in when the rule that it's learning may have our own biases encoded into it. For example, people who read résumés make decisions about which ones look like strong résumés and which ones don't. We wouldn't know how to write down a step-by-step procedure for that, but we can feed the results to a machine learning-styled algorithm. The algorithm will now try to learn a rule that distinguishes the résumés that look strong from the other ones.

This is where bias sneaks in. We have several decades of research from behavioral sciences that when people look at a résumé, a huge amount of their own implicit bias comes into the process. The algorithm—which knows nothing about the world—now just knows, "These are the strong résumés and these aren't." It's just trying to tell you a rule that faithfully describes your behavior, but your behavior was biased. It finds exactly the ways in which you're biased, and it reproduces them.

So, we think that we are removing the human element, perhaps, by using a machine learning algorithm. But that algorithm is learning things that we didn't even know about ourselves—and formalizing them.

Right. In addition to thinking of the algorithm as producing a tool, it is also producing a diagnostic.

The algorithm almost becomes like an experiment, which I can probe. I can create synthetic job applicants. I can run them through the algorithm. I can say, "Okay, what if I change it slightly this way? What happens?"

With a human being, if I ask them, “Would you still have hired this person if they had gone to school X instead of school Y?” the person might make their best effort to give you an answer to that. But they can’t really know what they would’ve done in that situation. With an algorithm, we can change the input from Y to X—what school the applicant went to, for instance—and we can just feed it back through. And we’ll learn something.

There is a sense in which we have a much better chance of understanding the pipeline of decision-making when it’s passed to an algorithm than when it’s a human being.

Complexity, transparency, and democracy

You recently authored a paper with one of your computer science grad students looking at how the government could use algorithms to more fairly and efficiently allocate stimulus checks. That’s a very timely question, and I will freely admit that—not being in the field—I could not begin to understand the findings!

The complexity of it made me wonder about how we bridge the gap between making smarter, more efficient policy decisions, and still having the voting public understand and have faith in what’s going on. How do you see that balance being achieved?

Those are great questions and big challenges. The question of making policy decisions that are informed by complex models and large amounts of data—that’s a problem that began before the widespread use of computing and algorithms, with the introduction of large-scale mathematical and statistical models into policymaking.

But the introduction of machine learning algorithms in computing takes us one step further because it allows us to deal with models that are, in some sense, inscrutable even to their developers. We can actually be looking at the answer, right in front of us—we have this computer code that is doing this thing that we don’t know how to do, and we can’t say how it’s doing it. It’s a profound challenge, and it’s still a very new area—the area of interpretability and explainability of machine learning algorithms.

Sticking with the example of stimulus checks: A simple solution everybody can understand is to give everyone the same amount of money—maybe subject to some basic rules and cutoffs. I believe your paper’s point is that complex algorithms could help target the assistance better, which would be a more efficient use of taxpayer money. But that runs into problems in a democracy. How can we make things more efficient, without just saying, “We just have to trust the robots to get it right”?

It’s a great example to work through because the first question you come to is one you can’t derive using an algorithm: What are we trying to achieve through the allocation of a stimulus? We could have some aggregate measure of economic activity that we’re trying to promote. We could try to maximize the number of people that we bring above some threshold that we’ve defined.

Those are human decisions that, in a democratic society, the policy process has to actually arrive at a conclusion on. I almost think of the role of the algorithm, or of mathematical models, as a counterparty in a dialogue about how to set objectives, how to set thresholds.

Lindsay France/Cornell University

You go to your model and you run a counterfactual simulation. You say, “What if we tried this, what would happen?” And then you see what happens, at least within your model. There’s this back-and-forth dialogue, where, in a sense, the computational model is giving you some clarity on the downstream consequences of choices that you might make.

The algorithm is telling us things that are very, very hard to figure out. Like, when I allocate [financial] assistance to a particular part of the system, everything is connected in some kind of network of transactions. It’s like I poke this spiderweb of transactions and the whole thing ripples and it spreads out in all sorts of different directions. Algorithms are very good at helping you figure out what all those ripples will look like. But then it’s up to you to decide what it is you’re actually trying to accomplish.

And going through that process of querying the model, calibrating it—that itself is potentially a way to build public, transparent faith that you are meeting whatever goals society sets out.

That’s how we hope the process works. The algorithm’s computational models are one participant in that process—and it’s a process with many participants.

In research with economist Sendhil Mullainathan, you make the point that when it comes to algorithms, simplicity and fairness can be fundamentally inconsistent with one another. That sounds like a very frustrating finding. What are we supposed to do with that knowledge?

We know that, as humans, if we’re operating under conditions of low information or rapid decision-making, that is when people are prone to fall back on stereotypes—and often pernicious stereotypes that work to the detriment of people who are already at a disadvantage.

If we take a complex model—let’s say there are thousands of pieces of information we might have about a person, and we could simplify it by using only a few pieces of information. What we found in this work was that when you start removing the information available to an algorithm, it begins to do things that resemble the human process of falling back on stereotypes.

What this tells us is that we should be alert to opportunities to strategically “un-simplify” our models in certain targeted ways. Th ere are many reasons to prefer simple models: [More complex algorithms] are inscrutable; they are not really amenable to collaborative decision-making or refining. But the question sometimes is: Are there ways in which, in a limited, targeted way, we can expand the models in ways that deliberately address the dimensions where it seems to be falling back on stereotype-like heuristics?

The argument for diversity … of algorithms

Here's another term for us—it's kind of a mouthful: "algorithmic monoculture." What is that, and what is the danger it can pose?

The term monoculture comes from agriculture, where if you plant the same plant species across all of your fields, it's at risk to being eliminated by a single pathogen that can sweep through the whole thing, or by a single change in weather conditions.

Suppose that we begin introducing algorithms for some problem that is very complicated and that humans struggle with: medical diagnosis, evaluating loan applications, evaluating résumés. Maybe we could even demonstrate that we have made the system more accurate, or we have reduced the amount of bias or disparity in the system.

We're now in a new kind of situation that becomes slightly precarious. Let's say all the different firms in an area are all doing a first-pass screening of résumés using the same algorithm. First, if the algorithm just doesn't like your résumé for some reason, you no longer have a chance for recourse or a second opinion. If one doesn't like you, then they're all not going to like you.

Second, if conditions change, then we could all suddenly start making the same set of mistakes. For example, maybe this is an algorithm that's evaluating loan applications, and the underlying economic conditions change. Maybe this model was trained pre-pandemic, when the meanings of certain things in your financial history just look different. Then, all of a sudden, all of these algorithms are now making mistakes in the same way because they're operating in an environment that they weren't trained on.

These are things that become much more acute risks now that we have the ability to really replicate our decisions through computing.

Computer science meets economics

You have research examining the power of algorithms to improve the way that we distribute welfare payments, or to improve intergenerational mobility. You get into one of the staples of behavioral economics, looking at sunk cost bias. As a computer scientist, what are you and your field bringing to these economic and social questions—and to the Institute?

I've gotten a huge amount of benefit, over my whole career, from working with economists and social scientists. What struck me through all of this collaboration is how many of the complex problems that we're dealing with involve systems that sit at the boundary of computational, economic, and social concerns. We're increasingly creating systems where people come together and they interact, and that interaction is often mediated by algorithms.

The interface by which we engage with each other in commerce, for example, or exchange information on something like a social media platform—all of these have algorithms as intermediaries. For people thinking about human behavior and human society, this role of algorithms as mediators of so much of our activity means that you really have to take into account what these algorithms are doing.

Conversely, the design of algorithms is going to need to take into account the ways in which human beings are going to interact with the algorithms. You bring up the example of, say, sunk cost bias, or similar examples like time-inconsistency or present-bias—this interplay between humans with all of their behavioral biases and the algorithms that they interact with has really become a very rich topic for questions.

The allocation of resources in financial systems, the dynamics of the labor market, the ways in which policy decisions get arrived at through a synthesis of viewpoints from many different stakeholders—I think all of these are places where there's a productive interaction to be had between economics and the social sciences and computing.

Social media and seeing the matrix

You've been at this long enough that we can track your career alongside the growth of social networks. We can go back to 2006, when you gave presentations speculating about whether people were becoming too exposed on MySpace. You were talking at least 10 years ago about the implications of personalized news feeds driven by algorithms and our "filter bubbles." Do you feel like you had tried to warn us all about the risks of social media, and should we have done something different at some point?

I didn't think of myself as trying to warn people. When I or my students or co-authors gave talks about this, I think we were trying to draw attention to social media as a topic and saying, "This is serious. This has the potential to have a major impact on society." In 2006, this was a bit of a hard sell because social media was this sort of frivolous activity where we went online and we shared photos, and we talked about things in our social media lives, while meanwhile our everyday lives went their own way.

Heather Ainsworth for Minneapolis Fed

I think the argument that we tried to make in 2006—and in 2010, and in 2012—is that there's really less and less daylight between our online lives and our sort of "real lives" in the offline world, that these are really merging. Similarly, that you eventually won't be able to separate social media from the political process. So, we should sort of treat this with the gravity that it needs.

Maybe that was a warning. On the other hand, I think it's very hard to predict how these systems turn out. And even now, I would say about social media what I say about algorithms: Algorithms are a tool, and tools can be used both constructively and destructively. They can do both a lot of good and a lot of bad.

We think about social media perhaps in enabling a group of people intent on causing harm to get together. And we say, "Why do we have this thing that lets people intent on causing harm to coordinate and form a group and operate more efficiently?" But that same social media platform also lets people who have a particular disease, who are not necessarily getting the help they need, to find other people online with that same condition and form patient support groups and discuss strategies and remedies that can be enormously helpful in their lives.

A powerful tool can be used for many different things. I think that's the sense in which we've been trying to approach this, even now, even after all that's happened. But yes, I think one of our messages all along is that, at minimum, this is going to seriously intersect with your life, it's going to seriously intersect with society, with the political process, and with many other things.

I'm thinking again about machine learning algorithms teaching us about ourselves, revealing things about ourselves that we didn't know. And then, as a society, interrogating that and changing course. We're perhaps going through that very painful process right now with social media—understanding the effects that it has on us, and then tying that back to our values.

I think that's absolutely the case. It has also accelerated something that I think that we've seen even prior to the internet, which is the way media in general can be a powerful mechanism for polarization, depending how it's used.

MORE FROM THIS ISSUE ›

One paper that I find fascinating in this domain is a paper by three economists (Matt Gentzkow, Jesse Shapiro, and Matt Taddy) looking at the evolution of partisan language in political speeches over two centuries. They looked at whether people from different political parties are essentially using different language when they address the same topic—"disjoint vocabularies." They asked the question: Given one minute of a politician speaking, how accurately can you predict which political party they're from? They plot this curve over 200 years, and the curve suddenly jumps up quite sharply—not in 2004 or 2005 with the introduction of social media, but in 1994, roughly around the consequential midterm elections during Bill Clinton's first presidential term. If you go back, you see that there were some very deliberate strategies taken with respect to the media, [and] the vocabularies of the two parties began to diverge. It's a reminder that it's been happening at all different points through our recent history, enabled by all different forms of media and communication technologies. Social media is the most recent step in that progression.

It's noteworthy that their paper uses the modern tools of machine learning to look back over that history, and that it's enabled us to actually go back and see things in our past— potentially more sharply than we saw them at the time.

I don't know how hard most of us think about the consequences when we play something on Netflix or follow our nose to a news article that Google serves up for us in our feed. On the other hand, I wonder if you feel a little like Neo in "The Matrix": Everywhere you look, you see lines of code manipulating us.

Right—it is something I think about. I think there are a few questions you should ask yourself when you encounter things. Why am I seeing this? Is this tailored to me? Who is learning from what I'm doing right now? Do different platforms owned by the same company mean that data you're generating here are informing decisions over there? Maybe there's some actual economic relationship between them, maybe there's a data-sharing relationship.

In elementary school, I'm sure both of us learned some basic things like the difference between nonfiction and fiction; between an objective viewpoint and a subjective viewpoint; between a primary source and a secondary source. There's a whole new set of things that we need to be learning today that are just that basic, which I think we're having to figure out as we go along: the difference between personalized and non-personalized content; between a page that was populated by a human author, versus something that was created by machine learning; between content that is basically fixed and static, and content that is being dynamically populated and changes each time you go back to the page.

All of these are phenomena of the internet, based on the algorithms that are powering these systems, that are as fundamental as those things we learned as kids about the difference between fiction and nonfiction, or subjective and objective.

For somebody who understands the capabilities of algorithms and the power of technology, for good or ill, you seem to retain a lot of faith that by aggressively using it—by iterating and listening to the feedback—we will make progress. The answer is not to retreat.

Yes. I certainly feel very keenly the difficulties and challenges that we face, the ways in which things can easily go wrong. But I do feel that to work in this area, it is important to believe that there is potential to bring about improvements and benefits. Not to think that naively, to be falsely optimistic, or to think that solutions here are easy—but to think that solutions are possible and that this is a goal worthy of all our energy and our creativity.

This interview has been edited for length and clarity.

This article is featured in the Spring 2022 issue of *For All*, the magazine of the Opportunity & Inclusive Growth Institute

Jeff Horwich

Senior Economics Writer

Jeff Horwich is the senior economics writer for the Minneapolis Fed. He has been an economic journalist with public radio, commissioned examiner for the Consumer Financial Protection Bureau, and director of policy and communications for the Minneapolis Public Housing Authority. He received his master’s degree in applied economics from the University of Minnesota.