The Volokh Conspiracy

Mostly law professors | Sometimes contrarian | Often libertarian | Always independent

About The Volokh Conspiracy

Stealth Quotas

The Dangerous Cure for "AI bias"

STEWART BAKER | 10.10.2022 6:07 PM

You probably haven't given much thought recently to the wisdom of racial and gender quotas that allocate jobs and other benefits to racial and gender groups based on their proportion of the population. That debate is pretty much over. Google tells us that discussion of racial quotas peaked in 1980 and has been declining ever since. While still popular with some on the left, they have been largely rejected by the country as a whole. Most recently, in 2019 and 2020, deep blue California voted to keep in place a ban on race and gender preferences. So did equally left-leaning Washington state.

So you might be surprised to hear that quotas are likely to show up everywhere in the next ten years, thanks to a growing enthusiasm for regulating technology – and a large contingent of Republican legislators. That, at least, is the conclusion I've drawn from watching the movement to find and eradicate what's variously described as algorithmic discrimination or AI bias.

Claims that machine learning algorithms disadvantage women and minorities are commonplace today. So much so that even centrist policymakers agree on the need to remedy that bias. It turns out, though, that the debate over algorithmic bias has been framed so that the only possible remedy is widespread imposition of quotas on algorithms and the job and benefit decisions they make.

To see this phenomenon in action, look no further than two very recent efforts to address AI bias. The first is contained in a privacy bill, the American Data Privacy and Protection Act (ADPPA). The ADPPA was embraced almost unanimously by Republicans as well as Democrats on the House energy and enactment of any privacy bill in a decade (its supporters hope to push it through in a lame-duck session). The second is part of the AI Bill of Rights released last week by the Biden White House.

Dubious claims of algorithmic bias are everywhere

I got interested in this issue when I began studying claims that algorithmic face recognition was rife with race and gender bias. That narrative has been pushed so relentlessly by academics and journalists that most people assume it must be true. In fact, I found, claims of algorithmic bias are largely outdated, false, or incomplete. They've nonetheless been sold relentlessly to the public. Tainted by charges of racism and sexism, the technology has been slow to deploy, at a cost to Americans of massive inconvenience, weaker security, and billions in wasted tax money – not to mention driving our biggest tech companies from the field and largely ceding it to Chinese and Russian competitors.

The attack on algorithmic bias in general may have even worse consequences. That's because, unlike other antidiscrimination measures, efforts to root out algorithmic bias lead almost inevitably to quotas, as I'll try to show in this article.

Race and gender quotas are at best controversial in this country. Most Americans recognize that there are large demographic disparities in our society, and they are willing to believe that discrimination has played a role in causing the differences. But addressing disparities with group remedies like quotas runs counter to a deep-seated belief that people are, and should be, judged as individuals. Put another way, given a choice between fairness to individuals and fairness on a group basis, Americans choose individual fairness. They condemn racism precisely for its refusal to treat people as individuals, and they resist remedies grounded in race or gender for the same reason.

The campaign against algorithmic bias seeks to overturn this consensus – and to do so largely by stealth. The ADPPA that so many Republicans embraced is a particularly instructive example. It begins modestly enough, echoing the common view that artificial intelligence algorithms need to be regulated. It requires an impact assessment to identify potential harms and a detailed description of how those harms have been mitigated. Chief among the harms to be mitigated is race and gender bias.

feature of proposals to regulate algorithms." The White House blueprint for an artificial intelligence bill of rights, for example, declares, "You should not face discrimination by algorithms and systems should be used and designed in an equitable way."

All roads lead to quotas

The problems begin when the supporters of these measures explain what they mean by discrimination. In the end, it always boils down to "differential" treatment of women and minorities. The White House defines discrimination as "unjustified different treatment or impacts disfavoring people based on their "race, color, ethnicity, [and] sex" among other characteristics. While the White House phrasing suggests that differential impacts on protected groups might sometimes be justified, no such justification is in fact allowed in its framework. Any disparities that could cause meaningful harm to a protected group, the document insists, "should be mitigated."

The ADPPA is even more blunt. It requires that, among the harms to be mitigated is any "disparate impact" an algorithm may have on a protected class – meaning any outcome where benefits don't flow to a protected class in proportion to its numbers in society. Put another way, first you calculate the number of jobs or benefits you think is fair to each group, and any algorithm that doesn't produce that number has a "disparate impact."

Neither the White House nor the ADPPA distinguish between correcting disparities caused directly by intentional and recent discrimination and disparities resulting from a mix of history and individual choices. Neither asks whether eliminating a particular disparity will work an injustice on individuals who did nothing to cause the disparity. The harm is simply the disparity, more or less by definition.

Defined that way, the harm can only be cured in one way. The disparity must be eliminated. For reasons I'll discuss in more detail shortly, it turns out that the disparity can only be eliminated by imposing quotas on the algorithm's outputs.

The sweep of this new quota mandate is breathtaking. The White House bill of rights would force the elimination of disparities "whenever automated systems can meaningfully impact the public's rights, opportunities, or access to critical needs" – i.e., everywhere it matters. The ADPPA in turn expressly mandates the healthcare, insurance, or credit opportunities.

And quotas will be imposed on behalf of a host of interest groups. The bill demands an end to disparities based on "race, color, religion, national origin, sex, or disability." The White House list is far longer; it would lead to quotas based on "race, color, ethnicity, sex (including pregnancy, childbirth, and related medical conditions, gender identity, intersex status, and sexual orientation), religion, age, national origin, disability, veteran status, genetic information, or any other classification protected by law."

Blame the machine and send it to reeducation camp

By now, you might be wondering why so many Republicans embraced this bill. The best explanation was probably offered years ago by Sen. Alan Simpson (R-WY): "We have two political parties in this country, the Stupid Party and the Evil Party. I belong to the Stupid Party." That would explain why GOP committee members didn't read this section of the bill, or didn't understand what they read.

To be fair, it helps to have a grasp of the peculiarities of machine learning algorithms. First, they are often uncannily accurate. In essence, machine learning exposes a neural network computer to massive amounts of data and then tells it what conclusion should be drawn from the data. If we want it to recognize tumors from a chest x-ray, we show it millions of x-rays, some with lots of tumors, some with barely detectable tumors, and some with no cancer at all. We tell the machine which x-rays belong to people who were diagnosed with lung cancer within six months. Gradually the machine begins to find not just the tumors that specialists find but subtle patterns, invisible to humans, that it has learned to associate with a future diagnosis of cancer. This oversimplified example illustrates how machines can learn to predict outcomes (such as which drugs are most likely to cure a disease, which websites best satisfy a given search term, and which borrowers are most likely to default) far better and more efficiently than humans.

Second, the machines that do this are famously unable to explain how they achieve such remarkable accuracy. This is frustrating and counterintuitive for those of us who work with the technology. But it remains the view of most experts I've consulted that the reasons for the algorithm's success cannot really be explained or understood; the machine can't tell us what subtle clues allow it to predict tumors from an apparently clear x-ray. We can only judge it by its

Still, those outcomes are often much better than any human can match, which is great, until they tell us things we don't want to hear, especially about racial and gender disparities in our society. I've tried to figure out why the claims of algorithmic bias have such power, and I suspect it's because machine learning seems to show a kind of eerie sentience.

It's almost human. If we met a human whose decisions consistently treated minorities or women worse than others, we'd expect him to explain himself. If he couldn't, we'd condemn him as a racist or a sexist and demand that he change his ways.

To view the algorithm that way, of course, is just anthropomorphism, or maybe misanthropomorphism. But this tendency shapes the public debate; academic and journalistic studies have no trouble condemning algorithms as racist or sexist simply because their output shows disparate outcomes for different groups. By that reductionist measure, of course, every algorithm that reflects the many demographic disparities in the real world is biased and must be remedied.

And just like that, curing AI bias means ignoring all the social and historical complexities and all the individual choices that have produced real-life disparities. When those disparities show up in the output of an algorithm, they must be swept away.

Not surprisingly, machine learning experts have found ways to do exactly that. Unfortunately, for the reasons already given, they can't unpack the algorithm and separate the illegitimate from the legitimate factors that go into its decisionmaking.

All they can do is send the machine to reeducation camp. They teach their algorithms to avoid disparate outcomes, either by training the algorithm on fictional data that portrays a "fair" world in which men and women all earn the same income and all neighborhoods have the same crime rate, or simply by penalizing the machine when it produces results that are accurate but lack the "right" demographics. Reared on race and gender quotas, the machine learns to reproduce them.

All this reeducating has a cost. The quotafied output is less accurate, perhaps much less accurate, than that of the original "biased" algorithm, though it will and gender constraints. To take one example, an Ivy League school that wanted to select a class for academic success could feed ten years' worth of college applications into the machine along with the grade point averages the applicants eventually achieved after they were admitted. The resulting algorithm would be very accurate at picking the students most likely to succeed academically. Real life also suggests that it would pick a disproportionately large number of Asian students and a disproportionately small number of other minorities.

The White House and the authors of the ADPPA would then demand that the designer reeducate the machine until it recommended fewer Asian students and more minority students. That change would have costs. The new student body would not be as academically successful as the earlier group, but thanks to the magic of machine learning, it would still accurately identify the highest achieving students within each demographic group. It would be the most scientific of quota systems.

That compromise in accuracy might well be a price the school is happy to pay. But the same cannot be said for the individuals who find themselves passed over solely because of their race. Reeducating the algorithm cannot satisfy the demands of individual fairness and group fairness at the same time.

How machine learning enables stealth quotas

But it can hide the unfairness. When algorithms are developed, all the machine learning, including the imposition of quotas, happens "upstream" from the institution that will eventually rely on it. The algorithm is educated and reeducated well before it is sold or deployed. So the scale and impact of the quotas it's been taught to impose will often be hidden from the user, who sees only the welcome "bias-free" outcomes and can't tell whether (or how much) the algorithm is sacrificing accuracy or individual fairness to achieve demographic parity.

In fact, for many corporate and government users, that's a feature, not a bug. Most large institutions support group over individual fairness; they are less interested in having the very best work force—or freshman class, or vaccine allocation system—than they are in avoiding discrimination charges. For these institutions, the fact that machine learning algorithms cannot explain themselves is a godsend. They get outcomes that avoid controversy, and they don't have to answer hard questions about how much individual fairness has been sacrificed.

will only know is that "the computer" found them wanting.

If it were otherwise, of course, those who got the short end of the stick might sue, arguing that it's illegal to deprive them of benefits based on their race or gender. To head off that prospect, the ADPPA bluntly denies them any right to complain. The bill expressly states that, while algorithmic discrimination is unlawful in most cases, it's perfectly legal if it's done "to prevent or mitigate unlawful discrimination" or for the purpose of "diversifying an applicant, participant, or customer pool." There is of course no preference that can't be justified using those two tools. They effectively immunize algorithmic quotas, and the big institutions that deploy them, from charges of discrimination.

If anything like that provision becomes law, "group fairness" quotas will spread across much of American society. Remember that the bill expressly mandates the elimination of disparate impacts in "housing, education, employment, healthcare, insurance, or credit opportunities." So if the Supreme Court this term rules that colleges may not use admissions standards that discriminate against Asians, in a world where the ADPPA is law, all the schools will have to do is switch to an appropriately reeducated admissions algorithm. Once laundered through an algorithm, racial preferences that otherwise break the law would be virtually immune from attack.

Even without a law, demanding that machine learning algorithms meet demographic quotas will have a massive impact. Machine learning algorithms are getting cheaper and better all the time. They are being used to speed many bureaucratic processes that allocate benefits, from handing out food stamps and setting vaccine priorities to deciding who gets a home mortgage, a donated kidney, or admission to college. As shown by the White House AI Bill of Rights, it is now conventional wisdom that algorithmic bias is everywhere and that designers and users have an obligation to stamp it out. Any algorithm that doesn't produce demographically balanced results is going to be challenged as biased, so for companies that offer algorithms the course of least resistance is to build the quotas in. Buyers of those algorithms will ask about bias and express relief when told that the algorithm has no disparate impact on protected groups. No one will give much thought (or even, if the ADPPA passes, a day in court) to individuals who lose a mortgage, a kidney, or a place at Harvard in the name of group justice.