Algorithmic Bias and Opportunity Risk

Saumi Rahnamay

Simon Fraser University

Phil 329: Law and Justice

Dr. Endre Begby

December 8, 2020

All over Canada and the United States, algorithmic processing is becoming more accessible and by extension more popular in certain governmental institutions. In this paper, I consider AI algorithmic decision-making used in assessing risk of recidivism in criminal courts, and AI used in police departments facilitating predictive policing. I argue that using algorithmic decision making is only an acceptable course of action if the risk associated in implementing that action is decisively low. I consider three different AIs: COMPAS, PredPol, and GeoDASH, and argue that all but PredPol are acceptable instances of algorithmic decision making. I begin by giving a quick summary of the state of algorithmic decision making in general, and recount ProPublica's article on COMPAS. Then, I offer a response to ProPublica originally fleshed out by Endre Begby. I then go on to further my own account of this response. I then consider the difference between assessing risk of recidivism, and predictive policing. And finally, I consider what this might say about recommendation trust in epistemic discussion and respond to Matthew Bennet's account of recommendation trust.

I. Brief Summary of AI in court rooms, and ProPublica's article on COMPAS

When someone is ruled guilty, courts must decide how severe the guilty party's sentence should be based on a variety of factors. One of these factors is the risk of the person committing crime again after their punishment is over; risk of recidivism. However, unassisted judgment from clinicians and other professionals are frequently inaccurate or biased. (Eichelman, 1982)

COMPAS and other similar AI algorithms aim to mitigate this issue using algorithmic processing to make better assessments of risk of recidivism. The rationale is, since unaided human judgment has proved to be more than fallible, and that tools (algorithmic or not) designed to increase accuracy in this measure do indeed help, algorithmic computing could be the most

powerful, and thereby most helpful tool in this endeavor. On the surface, computer decision making purports to offer a level of objectivity that cannot be achieved otherwise. The AI assesses risk of recidivism based on input criminal data and information about the convicted party, and calculates a score determining their level of risk. Higher risk scores influence courts in ruling longer/harsher sentences. It's important to note that COMPAS does not decide whether someone is guilty or not. Assessing guilt is a process handled by a judge and jury in court. Rather, once there is a guilty verdict, algorithms like COMPAS are one source that courts can draw upon in determining the severity of punishment.

COMPAS uses the same actuarial science methods used by insurance companies to determine risk of car accident, so as to determine at which rate to charge. Car companies look at data like the drivers age, gender, driving record, etc. Similarly, COMPAS is trained on reams of criminal profiles, and accounts for individual factors like criminal involvement, historical violence, financial status, social isolation, etc., which is then compiled into average criminal profiles for males and females of various age ranges. The offender's profile is fed to COMPAS which is then compared to a corresponding criminal profile. COMPAS is explicitly exclusionary of certain data sets. For instance: COMPAS is not told what race the offender is, in an attempt to bar any race-based assessments.

However, In ProPublica's *Machine Bias*, evidence is reproduced suggesting that COMPAS produces bias. The AI's risk assessment score tended to be higher for black people, and lower for white people. However, COMPAS' rate of failure (the rate at which it made wrong predictions) was much higher for black defendants than white ones. The machines assessments seem to be specifically discriminatory of black Americans. ProPublica coins a term for this kind of algorithmic prejudice: Algorithmic Bias. The idea is that the algorithm the AI is based on is

biased against black individuals. The purpose of implementing AI in the first place was to mitigate bias, and it seems that it's failed in doing so. Therefore, we should abandon algorithmic decision making because of this discrepancy.

II. Response to ProPublica – The machine isn't biased; society is.

Consider this response to ProPublica's claims, fleshed out by Endre Begby in an essay. (*see citation) ProPublica's conclusions rests on COMPAS' failure rate: although consistently scoring black people higher in risk than whites, these assessments were far more likely to be wrong than assessing white people. However, we may change our mind about COMPAS when considering what computations drive COMPAS to output these results. This disparity in failure rates turns out to be the result of a mathematical inevitability given unequal base rates (the number of white and black convicts). It turns out that, although failure rates were disparate, COMPAS' *success rates* were about equal between the two racial groups. This doesn't vindicate COMPAS from bias on its own; but in conjunction with unequal *base rates* between black and white defendants, higher failure rates are unavoidable. Given that COMPAS aims to maximize predictive success, and because black people make up a disproportionate number of defendants, it must follow that black people will be overrepresented in failure rates as well. (Kleinberg, et al. 2016)

The issue here is that the data being fed to the algorithm is itself biased against black people, and COMPAS notices that pattern through proxies. Although there are never any explicit mentions of race, race is often associated with other parameters that COMPAS does use. For example: since neighborhood race demographics are usually homogenous, an individual's zip code or address may be telling of their race. This is what Gabbirelle M. Johnson calls the proxy

problem. Reliable proxies of race can effectively side-step our attempts to mitigate race-based decisions.

The question now is directed to the existence of unequal base rates. Why is there a substantially disproportionate percentage of black defendants? (U.S. Department of Justice, 2011) What are the causal factors creating this asymmetrical distribution of crime in Canada and America? It seems that there are two possible domains these biasing factors can be found in. Bias may be at the institutional level; the law, police or any other national systematized mechanism may be constructed in such a way that inevitably disadvantages certain racial groups. Or bias may be found at the individual level; employers, citizens and state representatives may be racist and discriminatory towards certain racial groups, excluding them for public cooperation. There are certainly widely accepted candidates of what causes this disparity, such as poverty, segregation, or racism; each of which would require time and effort to rectify. There's no doubt in my mind that this civil injustice will be resolved in the generations to come. However, assessing risk of recidivism is not something we can put on pause until our data is better. Decisions need to be made; the show must go on. *For the time being, should we listen to COMPAS?* To answer this, we have to understand why our crime data is so lop-sided.

III. *Why the disparity?*

COMPAS is trained on past data of individual criminals, which it creates criminal models with, and goes on to compare newly convicted parties with. The bias is found somewhere along the process of creating that criminal model. As we've seen, it's not in COMPAS, it's in the data it's trained on; there's a disproportionate number of black offenders compared to their population. To investigate statistical bias, it is best to examine the conditions in which data is

gathered and work our way up to the results to see what goes wrong. In the following section, I argue there are at least two salient factors causing the statistical bias of black offenders.

A. Poverty is the mother of crime

Criminal activity is primarily intervened by police; a large portion of criminal data is gathered thereby. Police arrest criminals, who then record those events in databases. However, police do not spread themselves out evenly over their entire jurisdiction. Much like how crime is more likely to transpire at certain times (the median hour of crime), crime is more likely to transpire at certain locations as well. But since residential areas are typically homogenous in level-of-income (subsidized houses are typically built as apartments or neighborhoods to, presumably to cut costs), location serves as a reliable proxy for class. Indeed, the police bank on this proxy; data suggests citizens living in low-income households are more likely to be the victims of violent crime than higher-income households. (Statistic Canada, 2004) But it's also the case that annual wages for black men are among the lowest for any race; thus, low-income residency can be a proxy for a large population of black men.

This in itself is one factor for the disproportionate number of black offenders. Black men tend to find themselves in poverty, and so are more disposed to committing crime than other demographics. Although unjust, this bias is what we could call 'natural'; an uneven distribution of wealth among the races does exist in Canada and the United States; and thereby, as is the case of the distribution of crime. If this was all that was wrong with our data set, then we'd have a much harder time rejecting COMPAS. Let's consider another source of bias in the data.

B. Over-policing

Since more crime occurs in low-income communities, police tend to circulate those communities more than high-income communities. But as we've stated, low-income residency may also serve as proxy for the population of certain races in those residencies, which leads to a disproportionate level of policing for certain races. When a race is policed more than others, this creates sample bias in the crime data meant to represent the nation at large. Because there are only so many police officers to go around, disproportionate representation of one race takes away representation of another. The problem isn't that the crime data is misrepresentative of black people; more policing means more accurate reports of crime. The problem is that other races are comparatively *underrepresented* in the data, giving the impression that black people commit larger percentages of crime than others. The data would be different (maybe not vastly different) if police were evenly distributed among the races. Furthermore, police-stop report reviews suggest long-standing and widespread reports of discriminatory practices. Police were about six times more likely to target black individuals than white individuals for street checks in Halifax, a 2019 report describes. (Wortley, 2019) This suggests a level of implicit bias in policing, which undoubtedly has an effect on the crime data.

The overall bias in the crime data seems to exist as a combination of a couple factors: black men tend to find themselves in low-income communities and are also disproportionately over-represented and over-policed as compared to other races. It's entirely likely that there are other nuanced factors contributing to the disparity in crime data but explaining these two is enough for my purposes of discussing algorithmic decision making with 'dirty data'. Also, it's worth mentioning that, although I use black people in Canada and the United states as a prime example in my explanation, a parallel case could be made for indigenous individuals as well. (Browne, 2018)

Now that we understand the nature of bias in our crime data, we can decide whether to employ AI algorithms that use this data to aid in our decision making or not. In summary, the sources of bias we've here discussed are: because of socioeconomic circumstances, black men tend to resort to crime more than other races and are thus represented in crime data. And, because of targeted over-policing of low-income areas, and discriminatory over-policing of black people, black people are much more finely represented in criminal data, as opposed to other races.

IV. *Appropriate algorithmic decision making*

As mentioned in the previous section, race-crime statistics are disproportionately representative of black individuals, and unrepresentative of other races. The instances of crimes committed by black individuals is not inaccurate, but the proportion of crimes committed distributed by race is. Acknowledging this fact can help us decide when and when not to use AI decision making.

Since we have more data of black people than other races in crime, it seems that COMPAS is not prejudiced against blacks in its judgement. Even accepting that black individuals are overrepresented, that doesn't mean that the data used to train COMPAS is inaccurate. In fact, COMPAS' criminal profiles of black individuals may be more accurate than other races because of this abundance of data. We can now see that the issue isn't that black people are scored too high in risk; it's that other races are scored too low. Given this fact, courts may also understand that the more the offender's race is underrepresented in criminal data, the lower COMPAS will score. Courts may take this into consideration when consider COMPAS in assessing risk of recidivism in order to mitigate bias. So, now that we know specifically where COMPAS fails, courts can extend measures to mitigate those failures, thus increasing accuracy of assessing risk of recidivism. The risk of using COMPAS may be lower than we once thought,

so for that reason we ought to embrace algorithmic decision making in this setting. Furthermore, the alternatives to algorithmic decision making are antiquated and less accurate surveying techniques, subject to human bias. In the interest of remaining accurate in our predictions, COMPAS might be our best bet as of now. That being said, if we had a lot to risk in this situation by implementing COMPAS, then I would strongly advise being safer than sorry. But since COMPAS doesn't risk making the situation much worse than it already does, the opportunity-risk estimate is acceptable.

I'm not arguing that AI's should make decisions for us. Much like how we understand the fallibility of human judgment, we should not have the impression that algorithmic decision-making is objective in any sense. The AI's output can be viewed as a type of testimony not dissimilar to our own in its capacity to be wrong by happenstance, and so should be consulted similarly in court. Conversely, there are certainly instances where it's too risky to implement algorithmic decision making. When the situation is risky, and our data is dirty, algorithmic decision making imposes a hefty opportunity risk. In these scenarios, it's not only that algorithms get it wrong sometimes, like we do; they have the potential to make things substantially worse than they already were.

The risky case we will be examining is in policing. All over Canada, and the United States, some local police departments have adopted AI technology to aid in their endeavors to clean up crime. AI algorithms are fed various datasets and creates predictions of where and when criminal activity is likely to transpire in the form of heat maps. Law enforcement may be dispatched to the most criminally active areas; stopping crime before it happens. On paper, predictive policing also seems like algorithmic processing being used for a noble cause.

PredPol is one of such AIs. PredPol has been used in the united states by local police departments. However, it has been criticized for its disproportionate effect on minority communities. (Castelvecchi, 2019; Haskins, 2019) The data PredPol uses are from police data banks, like COMPAS. GeoDASH is another predictive policing software used in Vancouver, BC by the VPD, which is rife with similar issues. However, the VPD have a particular stance on predictive policing technology, which will be of interest to us later.

One might be tempted here to criticize predictive police technology of 'self exciting'. The algorithm sends police to somewhere it predicts crime will be,- whereby more crime will be recorded in our data by police officers so as to be fed back into PredPol or GeoDASH, creating a loop of prediction and data recording. However, creators of PredPol argue the software doesn't use crime types that have the possibility of being biased by police officers. It's only trained on crimes reported by victims, such as burglaries and robberies. (Castelvecchi, 2019) GeoDASH also uses this type of data. (Khoo, 2019) Therefore, PredPol and GeoDASH cannot be accused self-exciting in this way. However, there are other points of criticism.

Predictive policing is a risky endeavor because it has the potential to make matters much worse than they already were. Seemingly paradoxically, more policing in criminal areas doesn't seem to reduce crime as linearly as we'd initially suspect. (Di Tella et al. 2019) So, although PredPol and GeoDASH may be correctly predicting criminal activity, it might make matters worse if we doggedly followed these algorithms. This differs from the COMPAS case because the potential and level of disaster are higher. The opportunity-risk associated with policing is far too high to reasonably accept. For this reason, local police departments shouldn't use PredPol or GeoDASH. I will qualify my claim: police departments shouldn't use predictive policing technology *unless* they have the means to mitigate it's adverse and discriminatory side effects.

One police department that does have the capacity to mitigate predictive policing's adverse effects is in the VPD. The VPD is the second largest police force in British Columbia, after the RCMP. As aforementioned the VPD uses GeoDASH, a predictive policing algorithm locating where break-and-enter crimes are likely to occur for a 24 hour interval, in 2 hour periods (Vancouver Police Department, 2017). GeoDASH can be accessed online by the public, but only displays past reported crimes publicly, not predictions of criminal activity (geodash.vpd.ca). It uses the same type of data as PredPol: data arising exclusively from civilian reports to police. An interview with S/Constable Ryan Prox (Khoo, 2019), the VPD is aware of the potential for negative consequences in predictive policing, so they have taken steps to mitigate potential discrimination from GeoDASH. For example, a unique feature of GeoDASH is it's use of 'exclusionary zones', which account for areas where over policing concerns are present, such as the Downtown Eastside. The VPD also monitors how often GeoDASH deploys officers to any given neighborhood in Metro-Vancouver, especially those areas considered "socioeconomically or culturally sensitive", such as the South Slope, which has a large Indo-Canadian population. That's right, the VPD examines meta-data bias to implement exclusionary zones and other measure to mitigate criminal bias. The VPD also holds meetings every 3-4 months between the unit of police using predictive policing, and community police units, who operate more personally as tailored to their community. The point in dividing these units is so special attention can be effectively administered without higher orders from an AI. These meetings allow units to inform each other of potentially sensitive communities that has gone under the other's radar. For example, the VPD conferred with social housing initiative programs, and monitored overrepresentation of those areas by GeoDASH.

As we can see, the VPD takes extensive measures to mitigate adverse consequences of predictive policing. The VPD is privileged in their ability to do so, since they're such a large institutional branch. If the VPD didn't have the means to mitigate the high opportunity-risk of using predictive policing, it would be in the same category as PredPol. However, the VPD displays genuine concern of misusing predictive policing technology, and thus lowers the chance of potential disaster.

V. *Conclusion: Favorable Risk-Reward enables use of AI decision making.*

What makes the use of GeoDASH and COMPAS acceptable, and the use of PredPol unacceptable are varying levels of risk to reward in the cases we examined? Predictive policing does provide a great benefit to police institutions, and when used correctly can make positive differences. However, these benefits need to be considered in relation to the risk of incurring detriment by implementing this technology. As demonstrated, GeoDASH as implemented in Vancouver and COMPAS do not impose an unreasonable risk to reward ratio, while PredPol did because of it's irresponsible use. If PredPol used the same mitigatory measures that GeoDASH does, then it would no longer impose a serious risk. However, since many police departments cannot extend such measures due to budget or staff shortage, implementing algorithmic decision making would be detrimental. We should only implement algorithmic decision making in low risk settings.

VI. *AI and Recommendation Trust*

In his recent paper *Should I do as I'm Told? Trust, Experts, and COVID-19*, Matthew Bennett argues that following expert advice demands more trust than simply believing experts. To do this, Bennet distinguishes between recommendation trust; the level or type of trust necessary to adhere to someone's recommendations, and epistemic trust; the level or type of trust necessary to believe someone's testimony. He argues based on this distinction that increases in epistemic trust does not necessarily equate to increases in recommendation trust as well. Increasing public trust in expert recommendation is more demanding than simply improving the epistemic relations between the public and experts, especially amidst the COVID-19 pandemic. However, I will express why I believe Bennett's distinction between the two types of trust is ill-conceived and will expand my own account of recommendation trust based on the principles we've discussed above in algorithmic decision making.

To argue that increasing public trust in expert judgements may not cultivate the trust in recommendations based on those judgments, he considers the effects of implementing measures for greater transparency of value judgments made in the process of experiment and recommendation. (254) For example: Epidemiological models inform experts to create good recommendations to the nation. These models are only accurate if they account for hypothetical policies implemented by the government which would create different outcomes. Policies are based on value judgments. (ex: valuing economic security over lives saved is associated with a certain policy). Increasing transparency of the hypothetical values implemented in epidemiological research would set a precedent of honesty and thereby increase epistemic trust.

Bennet argues that implementing transparency might not be helpful in cultivating recommendation trust. And thereby, Bennet concludes that increases in epistemic trust does not

necessarily entail increases in recommendation trust. But, as Bennet implies earlier in his paper (see page 250), recommendations are hypothetical imperatives: If you desire X, you ought to do Y, because Y is the best way for X to obtain. In this light, transparency measures in any context would necessarily establish recommendation trust, because it proves that Y is the best course of action for X to obtain. Indeed, If I didn't (epistemically) trust you that Y was the best course of action for X to obtain, I wouldn't trust your recommendation. But the more justification you give, the higher my credence of that belief goes. Inch, by inch, I become more and more confident that your claim is true. At a certain level of credence, I concede, and trust your recommendation.

All this is to suggest that epistemic trust comes in degrees, levels of belief, certainty, credence, confidence, etc. I only trust your recommendation once my degree of belief in your claim has reached a sufficiently high level. There is a certain threshold of epistemic trust that needs to be crossed for there to be recommendation trust. However, I anticipate that epistemic trust is only an important factor because of it's inverse correlation with *perceived risk*. By increasing my knowledge of a state of affairs I rule out certain negative possibilities, thereby decreasing risk and begetting trust. The reason I can't trust your recommendation without justification is because of a high level of perceived risk. In other words, *it's not when our epistemic trust reaches a threshold for there to be recommendation trust; it's when our perceived risk of failure falls beyond a certain threshold.* The threshold for recommendation trust seems to be higher, or lower based on how high the stakes are for any given scenario.

This is precisely why AI recommendation can be very helpful, and trustworthy in certain situations, but unacceptable in other situations. Predictive policing is higher stakes than assessing risk of recidivism because it has a much higher potential for detriment. This is what makes

GeoDASH acceptable as implemented in Vancouver, while PredPol was openly criticized. GeoDASH takes steps to lower the stakes, thus lowering the necessary levels of epistemic trust. Furthermore, this is why at first glance, artificial intelligence seems trustworthy. Since their predictions are touted to be objective in some sense, their recommendations would be decisively low in risk. But, as the platitude goes “Garbage in, Garbage out”: our AI’s recommendations are only as trustworthy as our data is.

Bibliography

Eichelman, The clinical prediction of violent behavior, by John Monahan. Washington, DC: US Government Printing Office. DHSS Publication No. (ADM) 81–921,1981,134 pp, 1982

Northpointe, Practitioners Guide to COMPAS http://www.northpointeinc.com/files/technical_documents/FieldGuide2_081412.pdf

*Endre Begby, Automated Risk Assessment in the Criminal Justice System: A Case of 'Algorithmic Bias'? (Forthecoming)

Jon Kleinberg, Sendhil Mullainathan, Manish Raghavan, Inherent Trade-Offs in the Fair Determination of Risk Scores, 2016, arXiv:1609.05807

U.S. Department of Justice Office of Justice Programs Bureau of Justice Statistics, Alexia Cooper and Erica L. Smith, BJS Statisticians, Homicide Trends in the United States, 1980-2008

Statistics Canada, General Social Survey, Canadians from low income households experience higher rates of violent victimization, 2004.

Scot Wortley, Halifax, Nova Scotia: Street Checks Report (March 2019) at 105 <https://humanrights.novascotia.ca/streetchecks>

Rafael Di Tella and Ernesto Schargrodsky, Do Police Reduce Crime? Estimates Using the Allocation of Police Forces after a Terrorist Attack, The American Economic Review, Vol. 94, No. 1 (Mar., 2004), pp. 115-133

Davide Castelvecchi, Mathematicians urge colleagues to boycott police work in wake of killings, Nature, https://www.nature.com/articles/d41586-020-01874-9

Caroline Haskins, Academics Confirm Major Predictive Policing Algorithm is Fundamentally Flawed Vancouver Police Department, “Vancouver Police adopt new technology to predict property crime” (21 July 2017) .

https://geodash.vpd.ca

Interview of Ryan Prox by Cynthia Khoo & Yolanda Song (7 May 2019).

Matthew Bennett, Should I Do as I’m Told? Trust, Experts, and COVID-19, 2020