The Challenges of Algorithmic Bias

Jacquelyn Burkell, PhD
Associate Professor, Faculty of Information and Media Studies
The University of Western Ontario

Algorithms are everywhere. Chances are that before you got out of bed this morning an algorithm, running on your phone, was ready to offer the optimal route to the location that it determined – algorithmically – was your likely destination. Netflix, iTunes, and Amazon are using algorithms to recommend your media choices; Google is using an algorithm to rank your search results and shape your news feed. If you apply for a mortgage, an algorithm is determining your credit score; another algorithm is selecting the credit rate offer that you will receive in the mail for that new card you do not need but might apply for anyway – and when you receive that offer, rest assured that it is because an algorithm has already determined that *this* is the offer most likely to entice you to apply.

These, and many others, are examples of *predictive* algorithms – algorithms that make guesses about the future in order to make recommendations or decisions. Your phone analyzes your past movements and historic traffic patterns to recommend the best route to work – deciding on that destination, because that is where you typically go on a weekday. Media companies combine information about you with information about previous choices – yours and those of others like you – to identify items they predict you will enjoy (or at least select). Credit scoring leverages information about past borrowing and borrowers in order to predict default risk, and credit companies use the information to make loan decisions and set interest rates. Advertisers use details about information seeking, purchases, and demographics to deliver ads that are predicted to attract viewer attention.

Predictive algorithms abound in the justice system as well. Predictive policing uses algorithms to allocate policing resources to locations where, according to the algorithm, those resources are most likely to be needed (Perry, 2013). In the US, predictive algorithms determine the likelihood an offender will recidivate (Equivant, 2019) – information that is taken into account for bail and sentencing decisions. Some argue that AI might even, eventually, replace judicial decision making by analyzing past decisions and extrapolating the results for new decisions (Sourdin, 2018; but see Kerr and Mathen, 2014). In Estonia, this is already well on the way, with AI 'judges' in development that will make decisions in small claims court.[1]

Prediction is a natural, and indeed necessary, human endeavour (Schauer, 2009). Our survival as a species depended on our ability to predict whether an animal would attack or retreat, and where the best hunting might be found; today, individual survival depends on predictions such as the behaviour of drivers who share the road with us (is that car going to turn in front of me, or wait?).

[1] https://150sec.com/estonia-to-empower-ai-based-judge-in-small-claims-court/10985/

Top-down predictions influence and shape our perceptions (O'Callaghan et al., 2017); prediction is core to our cognitive capacities (Bubic et al., 2010); predicting what other people are going to do is central to social relations (Frith and Frith, 2006). Algorithmic prediction, then, is new only in that the mechanism of prediction is computational: prediction itself is a familiar feature of human decision making.

There is an inescapable ground truth about prediction, be it human or algorithmic: prediction of the future depends entirely on information about the past. Predictions are essentially projections of history into a future world. In order to make these predictions, we extract patterns or relationships from the complex, variable, messy, and incomplete information we have about the past. Our predictions rest on the assumption that regularities discoverable in the past will (or in the case of some decisions based on predictions, *should*) manifest in the future.

Why Algorithmic Prediction?

Human beings are astonishingly good at recognizing patterns and relationships. In fact, the capacity comes so easily to us that it is difficult to see the challenge. As children, we learn to map patterns of sound to words and meanings; we learn to relate patterns of light and colour, shape and depth to objects that we can recognize in the world. Expert poker players (St. Germain and Tennenbaum, 2011), and experts in other areas (Dreyfus, 1997), recognize and exploit winning patterns in complex games. Some researchers have shown that participants can intuitively evaluate logical syllogisms (Morsanyi and Handley, 2012), and expertise often manifests as an intuitive situational understanding (see, e.g., Effken, 2007). One way to interpret these intuitive conclusions is as examples of complex pattern recognition.

The next thing you notice, after marveling at the capacity of the human brain to extract these critical relationships, is that we get it wrong a lot of the time. Many of these errors arise from limitations in the amount of information we can process and the speed at which our brains operate (see Simon, 1990, on satisficing and bounded rationality). The results include the propensity to see patterns and relationships where none in fact exist (e.g., Chapman, 1967) in large part because we often focus on only a subset of relevant information when looking for patterns or relationships (e.g., Tversky and Kahneman, 1973). Other limitations become evident if you pay close attention: for example, we are more likely to recognize patterns or relationships we expect rather than those that surprise us, thus more likely to ignore those that do not match prior expectations (see, e.g., Kassin et al., 2013). Some types of error, such as visual illusions, are systematic and predictable; others are idiosyncratic, and may even be the result of prejudice and bias (Krueger, 1996).

These inaccuracies in our recognition of patterns or relationships become evident when we leverage pattern-recognizing capacities to make predictions and/or decisions. There is a demonstrated tendency, for example, for employers to prefer male candidates over equally-qualified female candidates for male-dominated jobs (Koch et al., 2015), and even young children hold gender-based stereotypes about occupational roles (Wilbourn and Kee, 2010).

Historically, mortgage lending in the US showed significant racial bias against black and Hispanic applicants (Myers, 1995). Similar issues occur in the justice system context. Judges demonstrate racial bias in bail decisions (Arnold et al., 2018); and police demonstrate a similar bias in vehicular stops (Warren et al., 2006). Male immigration applicants are disadvantaged when they appear in front of all-male panels (Gill et al., 2019). Indigenous children are over-represented in the children's aid system (Sinclair, 2016), at least in part due to stereotype-based assumptions about parenting capacity. In some cases, these biases are evidence of explicit prejudice and discrimination, but in many cases the bias is unrecognized and unintentional, possibly a manifestation of implicit bias (Jolls and Sunstein, 2006; Kassin et al., 2013) that operates below the level of consciousness and results in illusory correlations between variables that are in reality unrelated (e.g., Hamilton and Gifford, 1976; Smith and Alpert, 2007).

Automated evaluation systems seem tailor-made to address this difficulty. Computers are not subject to the same limits as human beings, and can process immensely amounts of information at incredible speeds – and they will only get better in this respect. If the problem with decision-making is that humans can't avoid use of the heuristics that lead to biased decisions (Tversky and Kahneman, 1974), then one answer is to remove the human factor, and have machines make the decisions and evaluations for us. That way the factors – both systematic and idiosyncratic -- that lead humans to make biased decisions can be removed. The algorithm, the reasoning goes, is *fair*.

Fair Algorithms?

In some ways and in some cases, this reasoning is absolutely correct. In particular, the use of algorithms to make decisions can ensure that idiosyncratic biases are removed from decision making (Cowgill, 2019). Unless explicitly designed to do so, an algorithm making a credit risk determination will not be swayed by personal connection; neither will an algorithm tend toward distrust as a result of an unrelated emotional state (Dunn and Schweitzer, 2005). Indeed, and importantly, algorithms have been demonstrated to make 'better' – fairer, more consistent, and more equitable – decisions than human decision-makers in a wide variety of application domains, including mortgage lending (Gates et al., 2002), selection of corporate board members (Erel et al., 2018), and employment decisions (Cowgill 2017; see also Kleinberg et al., 2017).

However, algorithmic decisions are neither always nor by all measures *fair*. Algorithmic bias, like human bias, abounds. In 2018, for example, Lam et al. found that Google image searches for many professions (including CEOs) under-represented women in the results; Nikon's 'blink' detector incorrectly identifies Asian faces as blinking in many images (Garcia, 2015); Facebook's algorithms for delivering targeted advertisements show gender bias (Ali et al., 2019); Equivalent's COMPAS algorithm for recidivism prediction shows, by some measures, racial bias (Angwin et al., 2016).

These demonstrations of 'unfair' algorithms raise a critical question: aside from intentional bias, which is outside the scope of this discussion, how does/can bias arise in automated prediction, and how can it be addressed?

Non-Representative Data

Predictive algorithms extract patterns over the data that is provided. If Nikon's 'blink' detector is trained on a dataset that under-represents Asian faces, the resulting algorithm cannot take the special characteristics of those faces into account when detecting blinks. If Google uses the images that are available online and tagged as 'CEO' as the source of image search results, and if women are under-represented in that group of images, women will be under-represented in the results. If a voice recognition algorithm is trained on male voices, it will under-perform on the voices of women. Predictive algorithms require a great deal of high-quality data, and data that are unrepresentative or poor quality will result in poor, and potentially biased, predictions. In general, ensuring that input data fairly and fully represent the population over which predictions are to be made will help to reduce algorithmic bias. The MIT Data Nutrition Project[2] is one of a number of initiatives to evaluate and certify the quality of the datasets that 'feed' algorithmic prediction, specifically machine learning applications. In general, those developing and those using algorithmic prediction should carefully consider the degree to which the input data reflect the full range of circumstances over which predictions are to be made.

In some cases, problems of representation can be more difficult to recognize. Consider an algorithm designed to assist graduate programs to select applicants who are likely to be successful. The data that are available over which the prediction can be made consist *only* of information about those applicants who have been accepted into the program (or into comparable programs). There is no information available about the outcomes for applicants who were *not* accepted into the program. In particular, if acceptance decisions are subject to some form of bias (e.g., age-based discrimination), the group subject to discrimination will be under-represented in the data. The algorithm could, over the available data, identify the pattern that distinguishes successful from unsuccessful *accepted* students; however, there is no way that the algorithm could take into account the characteristics that might have made an *unaccepted* applicant successful in the program. Unless the factors that predict success among the unaccepted group are consistent with those that predict success in the accepted group, the algorithm will be unable to identify potentially successful applicants among the group against which there has been historical discrimination.

One source of bias in algorithmic prediction is exclusion. In considering the quality of a dataset over which predictions are to be made, it is important to think about who or what is *not* included – and why. The patterns that are extracted, and that are used to make predictions, will not and cannot accurately or fairly represent what is absent from those initial data.

Choosing the Wrong Predictors

Surprisingly, one of the most obvious, and ultimately least effective, ways to reduce bias in algorithmic decision making is to remove information about protected characteristics – or

[2] https://datanutrition.media.mit.edu/

characteristics that identify members of protected groups -- from the dataset. If gender bias in employment decisions is a concern, for example, it is possible simply to eliminate gender from among the information that is known about applicants. On the surface, this would seem to effectively remove any possibility of gender bias: If the gender of applicants is not specified, the reasoning goes, that characteristic cannot contribute to any evaluation of candidates.

The problem is that there are many pieces of information about a job candidate that can, individually or collectively, signal gender as effectively as an explicit indication. Some of these might include the particulars about previous employment (since women are more likely to occupy some positions than others), job interruptions (since women are more likely to take pregnancy and childcare related leaves), or part-time employment (since women may be more likely to juggle family and work responsibilities). The point is that unless the correlated predictors are not also removed from the dataset,[3] gender could remain as a 'shadow' predictor, influencing outcomes almost as effectively as if it were explicitly included. The same argument applies to almost any categorization variable, including any characteristic that distinguishes individuals on a protected ground. Removing the variable does not remove the possibility of bias based on that characteristic; indeed, some argue that removing social category information can even exacerbate bias by making that bias more difficult to detect (Williams et al., 2018).

The inclusion of social category information among predictors does not necessarily signal bias, and removing that information is unlikely to eliminate, and may even in some senses exacerbate, bias in predictive algorithms, There are, however, other aspects of the set of variable included in the data that underly the predictions that are relevant to questions of bias. In the age of 'big data', information abounds, and it is easy to assume that more – and more varied – data is better for prediction. In considering the value, and potential bias, of a predictive algorithm, however, it is important to consider the nature of the data that enters into the predictions. The COMPAS algorithm, for example, includes 'leisure/boredom' as a potential predictor of recidivism. They provide reasonable theoretical motivation for including this variable, and document extant research that establishes a correlation between it and the outcome of recidivism (Equivant, 2019). One might question, however, the defensibility of including this factor in the algorithm that is predicting recidivism. Would it not be better to support offenders by offering leisure opportunities, rather than (potentially) penalizing those who do not have those opportunities available by refusing bail and/or increasing sentences on the basis of an increased predicted risk of recidivism? In general, the it is important to ensure that the variable that enter into the predictions meet standards of *validity* (see Raghavan et al., 2019 for a discussion of this issue in the context of algorithmic employment screening.

Predicting the Wrong Outcome

Issues of validity can arise with prediction inputs – and the same issues can arise with the predicted outcome. The COMPAS system is designed to predict re-offending – but there is no way to directly measure that outcome. Instead, the system predicts (based on previous cases),

[3] Or unless other statistical measures are taken to create predictors that are unrelated to gender.

the likelihood of *re-arrest*, using this as a proxy for re-offending. It is quickly evident, however, that re-arrest is not a perfect indicator of the commission of a new offense, in part because a new offense can go undetected and therefore would not result in an arrest. In experimental research, the challenge of translating a conceptual outcome to a measurable variable is termed 'operationalization', and it is widely recognized that a specific measure will often only partially and potentially inaccurately reflect a concept of interest (see Lehr and Ohm, 2017, for a discussion of this issue in the specific context of machine learning).

The use of re-arrest as a proxy for recidivism is not only inaccurate – it is a racially biased outcome indicator, since racialized individuals are subject to differential policing and charging practices. Racialized individuals are more likely to be subject to police stops (Gelman et al., 2007; Warren et al., 2006) and more likely to be subject to drug arrests (Mitchell et al., 2015). The use of predictive policing has been demonstrated to lead to the entrenchment of racialized policing (Jefferson, 2018), with the result that racialized neighbourhoods receive more policing resources, and thus individuals in these neighbourhoods are more subject to policy scrutiny. These differential practices are likely to result in a higher incidence of re-arrest for racialized offenders – an incidence that is based only partially on the outcome of interest, which is the commission of a new offense. Choosing the right outcome variable – one that meets standards of validity -- is critical for effective and meaningful prediction.

Biased world

Finally, we come down to the crux of the problem – the issue that, in the end, will in many cases defeat the goal of an 'unbiased' prediction. To understand this issue, it is critical first to come back to the very nature of prediction – looking backwards in order to predict forward. There is, in the end, no other choice: in order to make predictions about the future we must look to the past, and the only predictions we can make about the future will be based on the patterns and relationships that we, or the automated tools we develop, identify in what has come before.

Consider the problem of recidivism detection. If we agree for the moment that the exercise of predicting who is likely to reoffend is a valuable one, then we have no choice but to look to what we know about past offenders and their behaviour to predict future re-offending. Even if we address the problem that new charges are not an accurate measure of new offenses, if we correct for the bias created by increased surveillance of black offenders relative to white offenders, if we ensure that the pool of data over which predictions are made accurately represents the population of offenders -- we are still left with a problem.

Consider the range of characteristics, determined by prior research and theoretical models, that are measured by Equivant and used to construct their recidivism predictions. These include, among others, scales that assess criminal associates, substance abuse, financial problems/poverty, vocational/educational problems, family criminality, leisure/boredom, residential instability, and criminal social environment (Equivant, 2019). The theory that relates these factors to recidivism is well documented, and the relationships themselves seem logical: having few financial resources, relatively little education, associating with other criminals, living

in a criminal social environment all seem like factors that *would* increase the likelihood of re-offending.

They also seem like – indeed *are* – factors that would differentially affect racialized individuals. In the US, African Americans are over-represented in the criminal justice system (Blumstein, 2015). African American neighbourhoods tend, in general, to have lower average incomes than those in neighbourhood populated by Caucasians (Reardon et al., 2015). Black children in the US experience more adversity, including income disparity, than do white children (Slopen et al., 2015). Even well-off African American youth experience and educational gap relative to their Caucasian counterparts (Gosa et al., 2007). This is not the way the world *should* be – but it is the way the world *is*, and these systematic differences are clearly the outcome of longstanding historic biases and prejudice. In the end, the COMPAS algorithm for assessing recidivism likelihood might be biased against African Americans because the *world* is biased against that same group.

Those developing predictive algorithms have recognized this problem of 'bias in, bias out' (see Mayson, 2019, for an extended discussion), and various innovative technical approaches have been proposed to mitigate bias. Xu (2019) offers a technical overview of some suggested approaches. A careful analysis of the impact of proposed strategies is certainly warranted, and it is important to understand the impact of each approach on the bias exhibited in algorithmic prediction. Logic would suggest, however, that even the most innovative analytic approaches cannot undo the deepest systematic bias inherent in the world, and thus in the data that drives predictions, and ultimately in the predictions themselves.

Conclusion

In the end, we have to come back to patterns – the patterns, or relationships, that are inherent in our shared history and our shared practices. When all is said and done -- when the right input is selected, the right variables are identified, the most sophisticated analytic strategies are engaged – predictive algorithms will *still* exhibit bias, for the simple reason that those algorithms can *only* use the past to predict the future, and bias is a fact of our history. We can't pin the problem on algorithms or those who produce them, and algorithms won't get us out of the mess: because many forms of bias are already 'baked in' to the world we live in.

Does this mean that we abandon algorithmic prediction? No. Algorithmic predictions are in many cases no worse, and can often be better, than the human predictions they replace, both in terms of accuracy, and in terms of bias. Algorithmic predictions can eliminate idiosyncratic biases, including those based in explicit prejudice, from decision making. Algorithms can make predictions faster than humans, and they can use more data in making those predictions. Algorithmic prediction can reduce human workload, and increase human decision-making capacity. Moreover, increasingly sophisticated analytic techniques can go some way toward addressing bias, potentially assisting us to recognize and mitigate historical decision-making biases.

Do we unequivocally accept algorithmic prediction? *Emphatically*, no. There are by now many, and will be many more, instances of egregious algorithmic bias arising from intended discrimination, problems with case and variable selection, and development of the prediction algorithm. Moreover, algorithms, by virtue of their widespread application, can extend and ossify existing bias – so there are reasons to reject even those algorithms that are not subject to those addressable issues.

What, then, are we to do? Ultimately, there is a choice to be made, and it is incumbent on us – as legal professionals, as members of the judiciary, as academics, as developers, as regulators, and as citizens – to participate in that choice. In determining whether and when to allow algorithms input into critical decisions, we must actively consider the quality of the data, the quality of the prediction, the tradeoffs resulting from utilization, and the ethical acceptability of automated prediction (Lehr and Ohm, 2017). We must, in the justice system and elsewhere, carefully consider the algorithms that we use. We should consider ourselves empowered in the evaluation of those algorithms, able to ask questions of developers and question the results of processes we do not completely understand. We should, and must, ask for explanations – explanations that meet social and policy requirements (Zerilli et al., 2018) – of algorithms and algorithmic decision making. We must demand audits (Lepri et al., 2018), including fairness audits that explicitly test for bias against protected groups (Saleiro, 2018) of algorithmic decision tools. We must, in the end, be active participants in decisions about whether, and how, to automate decisions.

Finally, a 'different take' on the issue of patterns. Automated tools, based on complex multivariate analyses or increasingly sophisticated machine learning approaches, have the capacity to 'surface' patterns that might otherwise remain invisible. With respect to bias, this is both the Achilles heel and the undeniable power of algorithms. The risk is that indiscriminate use of algorithms can lock us in to pre-existing and subtle patterns of discrimination and bias. The promise, however, can be equally important: those same techniques can alert us to discriminatory and biased practice (Haijan et al., 2016), and they can serve as an audit on human decision processes (e.g., Leavy, 2018), or even allow intervention to eliminate those practices (Daugherty et al., 2019).

References

Ali, M., Sapiezynski, P., Bogen, M., Korolova, A., Mislove, A., & Rieke, A. (2019). Discrimination through optimization: How Facebook's ad delivery can lead to skewed outcomes. *arXiv preprint arXiv:1904.02095*.

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2016). How we analyzed the COMPAS recidivism algorithm.https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm, accessed October 27, 2019.

Arnold, D., Dobbie, W., & Yang, C. S. (2018). Racial bias in bail decisions. *The Quarterly Journal of Economics*, *133*(4), 1885-1932.

Blumstein, A. (2015). Racial disproportionality in prison. In *Race and social problems* (pp. 187-193). Springer, New York, NY.

Bubic, A., Von Cramon, D. Y., & Schubotz, R. I. (2010). Prediction, cognition and the brain. *Frontiers in human neuroscience*, *4*, 25.

Chapman, L. J. (1967). Illusory correlation in observational report. *Journal of Verbal Learning and Verbal Behavior*, *6*(1), 151-155.

Cowgill, B. (2017). Bias and Productivity in Humans and Algorithms: Theory and Evidence from Resumé Screening, Working Paper, 2017.

Cowgill, B. (2019) Bias and Productivity in Humans and Machines (August 6, 2019). Upjohn Institute Working Paper 19-309, 2019. Available at SSRN: https://ssrn.com/abstract=3433737 or http://dx.doi.org/10.2139/ssrn.3433737

Daugherty, P. R., Wilson, H. J., & Chowdhury, R. (2019). Using Artificial Intelligence to Promote Diversity. *MIT Sloan Management Review*, *60*(2), 1.

Dreyfus, H.L. 1997. "Intuitive, deliberative, and calculative conceptual schemes of expert performance". In *Naturalistic decision-making*, Edited by: Zsambok, C. and Klein, G. 17–28. Mahwah, NJ: Lawrence Erlbaum.

Dunn, J. R., & Schweitzer, M. E. (2005). Feeling and believing: the influence of emotion on trust. *Journal of personality and social psychology*, *88*(5), 736.

Effken, J. A. (2007). The informational basis for nursing intuition: philosophical underpinnings. *Nursing Philosophy*, *8*(3), 187-200.

Equivant (2019). Practitioners guide to COMPAS. April 4, 2019. Available online at http://www.equivant.com/wp-content/uploads/Practitioners-Guide-to-COMPAS-Core-040419.pdf.

Erel, I., Stern, L., Chenhao, T., & Weisbacj, M. S. (2018). Could machine learning help companies select better board directors? *Harvard Business Review,* April 9, 2018.

Frith, C. D., & Frith, U. (2006). How we predict what other people are going to do. *Brain research*, *1079*(1), 36-46.

Garcia, F. (2015). 8 Times Technology Proved to Be Racist; Modern Tech has a history of privileging whiteness. *NTR SCTN.* Retrieved October 27, 2019. Available from http://ntrsctn.com/science-tech/2015/11/racist-technology/

Gates, S. W., Perry, V. G., & Zorn, P. M. (2002). Automated underwriting in mortgage lending: good news for the underserved?. *Housing Policy Debate*, *13*(2), 369-391.

Gelman, A., Fagan, J., & Kiss, A. (2007). An analysis of the New York City police department's "stop-and-frisk" policy in the context of claims of racial bias. *Journal of the American Statistical Association*, *102*(479), 813-823.

Gill, R. D., Kagan, M., & Marouf, F. (2019). The impact of maleness on judicial decision making: masculinity, chivalry, and immigration appeals. *Politics, Groups, and Identities*, *7*(3), 509-528.

Gosa, T. L., & Alexander, K. L. (2007). Family (dis) advantage and the educational prospects of better off African American youth: How race still matters. *Teachers College Record*, *109*(2), 285-321.

Hajian, S., Bonchi, F., & Castillo, C. (2016, August). Algorithmic bias: From discrimination discovery to fairness-aware data mining. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 2125-2126). ACM.

Hamilton, D. L., & Gifford, R. K. (1976). Illusory correlation in interpersonal perception: A cognitive basis of stereotypic judgments. *Journal of Experimental Social Psychology*, *12*(4), 392-407.

Jefferson, B. J. (2018). Predictable policing: Predictive crime mapping and geographies of policing and race. *Annals of the American Association of Geographers*, *108*(1), 1-16.

Jolls, C., & Sunstein, C. R. (2006). The law of implicit bias. *Calif. L. Rev.*, *94*, 969.

Kassin, S. M., Dror, I. E., & Kukucka, J. (2013). The forensic confirmation bias: Problems, perspectives, and proposed solutions. *Journal of applied research in memory and cognition*, *2*(1), 42-52.

Kerr, I. R., & Mathen, C. (2014). Chief Justice John Roberts is a Robot. *University of Ottawa Working Paper*.

Kleinberg, J., Lakkaraju, H., Leskovec J., Ludwig, J. and Mullainathan, S. (2017). Human decisions and machine predictions. *The Quarterly Journal of Economics, 133* (1), 237–293.

Koch, A. J., D'Mello, S. D., & Sackett, P. R. (2015). A meta-analysis of gender stereotypes and bias in experimental simulations of employment decision making. *Journal of Applied Psychology, 100*(1), 128-161.

Krueger, J. (1996). Personal beliefs and cultural stereotypes about racial characteristics. *Journal of personality and Social Psychology*, *71*(3), 536.

Lam, O., Wojcik, S., Broderick, B., and Hughes, A. (2018). *Gender and Jobs in Online Image Searches*, PEW Research Centre (December 18, 2018), accessed from file:///Users/jacquelynburkell/Downloads/JobsGender_report_FINAL1.pdf, October 27, 2019.

Leavy, S. (2018). Uncovering gender bias in newspaper coverage of Irish politicians using machine learning. *Digital Scholarship in the Humanities*, *34*(1), 48-63.

Lehr, D., & Ohm, P. (2017). Playing with the data: what legal scholars should learn about machine learning. *UCDL Rev.*, *51*, 653.

Lepri, B., Oliver, N., Letouzé, E., Pentland, A., & Vinck, P. (2018). Fair, transparent, and accountable algorithmic decision-making processes. *Philosophy & Technology*, *31*(4), 611-627.

Mayson, S.G. (2019). Bias in, Bias out. *Yale Law J. (forthcoming).* Available online at https://papers.ssrn.com/sol3/papers.cfm? abstract_id=3257004, accessed October 27,2019.

Mitchell, O., & Caudy, M. S. (2015). Examining racial disparities in drug arrests. *Justice Quarterly*, *32*(2), 288-313.

Morsanyi, K., & Handley, S. J. (2012). Logic feels so good—I like it! Evidence for intuitive detection of logicality in syllogistic reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(3), 596.

Myers Jr, S. L. (1995). Racial discrimination in housing markets: Accounting for credit risk. *Social Science Quarterly*, 543-561.

O'Callaghan, C., Kveraga, K., Shine, J. M., Adams Jr, R. B., & Bar, M. (2017). Predictions penetrate perception: Converging insights from brain, behaviour and disorder. *Consciousness and Cognition*, *47*, 63-74.

Perry, W. L. (2013). *Predictive policing: The role of crime forecasting in law enforcement operations*. Rand Corporation.

Raghavan, M., Barocas, S., Kleinberg, J., & Levy, K. (2019). Mitigating Bias in Algorithmic Employment Screening: Evaluating Claims and Practices. *Available at SSRN 3408010*.

Reardon, S. F., Fox, L., & Townsend, J. (2015). Neighborhood income composition by household race and income, 1990–2009. *The Annals of the American Academy of Political and Social Science*, *660*(1), 78-97.

Saleiro, P., Kuester, B., Stevens, A., Anisfeld, A., Hinkson, L., London, J., & Ghani, R. (2018). Aequitas: A Bias and Fairness Audit Toolkit. *arXiv preprint arXiv:1811.05577*.

Schauer, F. F. (2009). *Profiles, probabilities, and stereotypes*. Harvard University Press.

Simon, H. A. (1990). Invariants of human behavior. *Annual review of psychology*, *41*(1), 1-20.

Sinclair, R. (2016). The Indigenous child removal system in Canada: An examination of legal decision-making and racial bias. *First Peoples Child & Family Review*, *11*(2), 8-18.

Slopen, N., Shonkoff, J. P., Albert, M. A., Yoshikawa, H., Jacobs, A., Stoltz, R., & Williams, D. R. (2016). Racial disparities in child adversity in the US: Interactions with family immigration history and income. *American Journal of Preventive Medicine*, *50*(1), 47-56.

Smith, M. R., & Alpert, G. P. (2007). Explaining police bias: A theory of social conditioning and illusory correlation. *Criminal justice and behavior*, *34*(10), 1262-1283.

Sourdin, T. (2018). Judge v. Robot: Artificial Intelligence and Judicial Decision-Making. *UNSWLJ*, *41*, 1114.

St. Germain, J., & Tenenbaum, G. (2011). Decision-making and thought processes among poker players. *High Ability Studies*, *22*(1), 3-17.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive psychology*, *5*(2), 207-232.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *science*, *185*(4157), 1124-1131.

Warren, P., Tomaskovic-Devey, D., Smith, W., Zingraff, M., & Mason, M. (2006). Driving while black: Bias processes and racial disparity in police stops. *Criminology*, *44*(3), 709-738.

Wilbourn, M. P., & Kee, D. W. (2010). Henry the nurse is a doctor too: Implicitly examining children's gender stereotypes for male and female occupational roles. *Sex Roles*, *62*(9-10), 670-683.

Williams, B. A., Brooks, C. F., & Shmargad, Y. (2018). How algorithms discriminate based on data they lack: Challenges, solutions, and policy implications. *Journal of Information Policy*, *8*, 78-115.

Xu, J. (2019). Algorithmic solutions to algorithmic bias: A technical guide. Available online at https://towardsdatascience.com/algorithmic-solutions-to-algorithmic-bias-aef59eaf6565 . Accessed October 27, 2019.

Zerilli, J., Knott, A., Maclaurin, J., & Gavaghan, C. (2018). Transparency in Algorithmic and Human Decision-Making: Is There a Double Standard?. *Philosophy & Technology*, 1-23.