A Brief Primer on Bias in Machine Learning and Algorithmic Decisions

Viren Dias, Sriganesh Lokanathan, Yudhanjaya Wijeratne
LIRNEasia, 12 Balcombe Place, Colombo, Sri Lanka
(viren, sriganesh, yudhanjaya)@lirneasia.net

LIRNEasia is a pro-poor, pro-market think tank whose mission is *catalyzing policy change through research to improve people's lives in the emerging Asia Pacific by facilitating their use of hard and soft infrastructures through the use of knowledge, information and technology*.

Contact: 12 Balcombe Place, Colombo 00800, Sri Lanka. +94 11 267 1160. info@lirneasia.net
www.lirneasia.net

This work was carried out with the aid of a grant from the International Development Research Centre (IDRC), Canada.

Introduction

It may surprise the non-computing world that, for all the talk of algorithms, computer science actually offers multiple definitions of what an algorithm is (Knuth, 1968; Markov, 1954; Minsky, 1967; Stone, 1972). Donald Knuth—the legendary mathematician, Turing Award winner and "father of the analysis of algorithms" (Karp, 1986)—once defined an algorithm as having these five characteristics:

- *Finiteness.* "An algorithm must always terminate after a finite number of steps."
- *Definiteness.* "Each step of an algorithm must be precisely defined; the actions to be carried out must be rigorously and unambiguously specified for each case."
- *Input.* "...quantities which are given to it initially before the algorithm begins. These inputs are taken from specified sets of objects."
- *Output.* "...quantities which have a specified relation to the inputs."
- *Effectiveness.* "... all of the operations to be performed in the algorithm must be sufficiently basic that they can in principle be done exactly and in a finite length of time by a man using paper and pencil" (Knuth, 1968).

A cursory glance at this criteria reveals the nature of algorithms of yesteryear. They were generally formulaic—a human was responsible for every piece of the algorithm; and therefore it was accepted that a human being could *explain* what an algorithm did, right down to being able to do the calculation themselves; and that there was nothing in said algorithm that hadn't been put there by its designer.

However, it was soon recognized that some problems were too complex to be solved deterministically. For instance, recognizing a face: it is difficult to precisely describe, element by element, what a face looks like. A more appropriate approach to solving these complex problems was to construct algorithms that can learn from data—machine learning.

With machine learning came the shedding of the old, explainable design process: instead, an architecture would, over many training cycles, self-adapt and develop its own paths to produce the kind of output or task fitness required. Even relatively old and primitive efforts (by computer science time) produced solutions that defied explanation: for example, Alan Thompson's (1996) experiment in trying to get circuits to design themselves produced circuits that did the job with approximately a third of the resources they were supposed, with components that sometimes weren't connected to each other, and most likely relied on magnetic flux in the circuit to work - something no human designer could have predicted. This is the wave of what Andrej Karpathy (2017)—director of AI at Tesla—calls "Software 2.0": an increasingly prevalent stack that offers superior functionality in some domains but with the caveat that the core algorithm itself is difficult to explicitly define or design.

Our observation is that sophisticated machine learning violates at least three of Knuth's principles: *finiteness, definiteness* and *effectiveness.* We are long since past the pencil-and-paper stage.

Naturally, this has led to many concerns in the development field, particularly when machine learning (often referred to offhand as AI) interacts with socio-legal systems. Key among these concerns is the subject of this paper:

What do we do about bias in algorithms we don't understand?

What is Algorithmic Bias in a Machine Learning World?

Researchers at the Brookings Institute, concerned with much the same issues that we are, broadly define bias as "outcomes which are systematically less favorable to individuals within a particular group and where there is no relevant difference between groups that justifies such harms" (Lee et al., 2019).

In our observation and work, machine learning presents us with three broad avenues by which such biases may manifest in systems:

Task-Fitness Bias

Machine learning systems (or "AI") are not one-size-fits-all entities. However, as Andrej Karpathy (2017) points out, the general "stack" is much more homogenous than the previous generation of hand-built algorithms. This nature makes it possible to adapt a system trained for one task to another, as long as they are broadly within the same domain: for example, Open AI's GPT-2, which gained much attention for its realistic, fanciful generation of news articles (Radford et al., 2019), can be quite easily retrained to write poetry instead (Wijeratne, 2019).

However, such task-fitness needs to be examined very seriously before implementation, as the recent PREDPOLL uproar shows (Haskins, 2019): an algorithm based on earthquake modelling should not be used to detect crime. Even within domains, vestiges of the old system will always remain; and therefore one potential source of bias is from a system introducing artefacts from a task to which it had previously been trained.

Data-Driven Bias

Data-driven bias is the centerpiece of most conversations, and the most easily understood. This occurs when the data fed into a machine learning system encapsulates bias found in human societies. It can materialize in the form of incorrectly correlating certain concepts with certain demographics; for instance, if we were to train an algorithm on what a physicist looks like using images of historical physicists, it would be biased towards picking men, because in the past, due to the shape and nature of patriarchal influence, physicists have been predominantly male.

One of the longest-running and most visible examples of this has indeed been in large systems like the Google search engine, where user interactions and flawed, pre-existing data form rich datasets that contains such biases (Noble, 2018), often resulting in racially charged search data. Facebook, for example, produces echo chambers (Hosanagar, 2016). Other biases are sometimes surreal: Microsoft's AI chatbot "Tay", for example, which learned from user interaction, started mimicking racist content sent to it by the community that interacted with it (Vincent, 2016), while a similar Microsoft chatbot "Rinna", that

interacts with a different community, started exhibiting the language of depression (Brown, 2016).

Google (2017) has suggested the following subdivisions for data-driven bias:

- *Interaction bias*. The users bias the algorithm based on the way they interact with it—"…like this recent game where people were asked to draw shoes for the computer. Most people drew [standard shoes]. So as more people interacted with the game, the computer didn't even recognize [high heels]."
- *Latent bias*. The algorithm incorrectly correlates concepts with certain demographics—"…for example, if you were training a computer on what a physicist looks like, and you're using pictures of past physicists, your algorithm will end up with a latent bias skewing towards men."
- *Selection bias*. The algorithm favours certain demographics at the cost of others—"…say you're training a model to recognize faces. Whether you grab images from the internet or your own photo library, are you making sure to select photos that represent everyone?"

However, we believe that this additional layer of categorization serves no purpose other than to illustrate the mechanisms by which the training dataset can be skewed; the overarching problems are common across all three subcategories and the categories are not mutually exclusive.

Overfitting

This occurs when the relentless pursuit of accuracy in training produces a machine learning model that performs flawlessly in a test environment, but poorly in the real world, introducing large biases in decision-making in a system that may seem perfect "on paper". This is referred to as overfitting in computer science and mathematics (Cawley & Talbot, 2010).

These are broad categories. Note that much more detailed taxonomies of bias exist (Mehrabi et al., 2019), but these overarching divisions are useful for discussion in non-technical settings.

What Happens When These Biases Affect Systems?

Such algorithms influence many aspects of our lives, from the news articles we read, the videos, shows and movies we watch to the people we interact with and the results of our applications for jobs, loans, education, etc. (Pasquale, 2015).

Kate Crawford (2017)—co-founder of the AI Now Institute—suggested that the negative impacts of algorithmic bias can be broadly categorized into allocative and representational harms. We find this a useful framework for examining the impacts of systems:

Allocative Harm

An allocative harm is when an algorithm withholds an opportunity or resource from a certain demographic. It is characterized as; immediate: the result of an instantaneous decision, discrete: easily quantifiable, and transactional: the result of a specific transaction. Subcategories of allocative harms include:

- *Stereotyping.* An oversimplified idea of something that incorrectly correlates certain concepts with certain demographics. For instance correlating a certain ethnicity with a higher socioeconomic status.
- *Recognition.* A certain demographic is unrecognised by the algorithm. For instance a facial recognition system that has difficulty recognising the faces of minority groups.
- *Denigration.* Assigning culturally offensive or inappropriate labels. For instance, labelling Muslims as terrorists.
- *Underrepresentation.* Underrepresenting a certain demographic. For instance, a Google image search of "doctor" yielding predominantly images of white males.
- *Ex-nomination.* When the majority demographic becomes accepted as the norm and the deviation of minority demographics from this norm becomes evident. For instance, a person from Sri Lanka is assumed to be of Sinhalese ethnicity and Buddhist religious association.

Representational Harm

A representational harm is when an algorithm reinforces the subordination of certain demographics along the lines of race, class, gender, etc. It is characterised as; long term: a process that affects attitudes and beliefs, difficult to formalise and diffuse: the consequences are vague and indirect, and cultural: the result of depictions of humans and society.

A prime example of representational harm was when Google Photos incorrectly labelled an image of a black couple as "gorillas" (Alciné, 2015). Without context, it may seem like a

harmless error. However, if the history of blacks being enslaved and considered as less than human for centuries is taken into account, it is evident how damaging such an error can be.

Evaluating the impact of algorithmic bias from a different perspective—minimising bias in algorithms can in turn curb bias in society by means of a feedback loop. Consider the hypothetical case of a young woman aspiring to become a computer programmer. Statistically, computer programmers are predominantly male, and a Google image search of "computer programmer" will reflect this skew. This may discourage the young woman from pursuing such a career path, thus reinforcing the stereotype that computer programmers are male. However, if the algorithm were to be adjusted to present a larger portion of female computer programmers than what is actually present in society, then the young woman may feel motivated to continue along that career trajectory, thus mitigating the stereotype in society.

What is Fair?

Regardless of the method used, addressing the matter of algorithmic bias involves the singular goal of making the algorithm fair. The predominant conversational thrust in development is to insist that biases be corrected and flawed algorithms retired or controlled.

However, this is a more complex ideal than one would suppose. The definition of what is "fair" is obscure, and there is more than one apt definition, which would change contingent on the lens by which a situation is viewed. One of the most useful lenses we have come across is Kleinberg et al.'s (2016) three notions of algorithmic fairness. Consider the following scenario:

- The purpose of the algorithm is to make a *binary prediction*: positive or negative.
- Each member of the population has an associated *feature vector*—a set of variables—according to which the prediction is made.
- Each member is assigned to one of two *groups*, with respect to which we would like the algorithm to be unbiased.
- Within each group, members are divided into *bins* based on their feature vectors.
- Each bin is labelled with a *score* that represents the probability of a positive outcome for each member in that bin.

Then the three notions of fairness were suggested as follows:

- *Calibration within groups.* For each group and each bin the expected number of members with a positive outcome should be proportional to the score assigned to that bin.
- *Balance for the positive class.* The average score of members with a positive outcome should be the same for each group.
- *Balance for the negative class.* The average score of members with a negative outcome should be the same for each group (Kleinberg et al., 2016).

Kleinberg et al. (2016) then continued to argue that no method can satisfy all three notions of fairness simultaneously, with the exception of highly constrained special cases:

- *Perfect prediction.* For each feature vector, we know for certain what the outcome is.
- *Equal base rates.* The two groups have the same fraction of members that have a positive outcome.

Furthermore, even satisfying all three notions approximately would require an approximate version of these special cases. It is helpful to illustrate these claims with a real world example:

Case Study: Northpointe's COMPAS Recidivism Algorithm

One of the most powerful case studies we can use is that of Northpointe: a Michigan-based company that developed an algorithm for predicting recidivism, named Correctional Offender Management Profiling for Alternative Sanctions (COMPAS). The algorithm was used to assist in making judicial decisions.

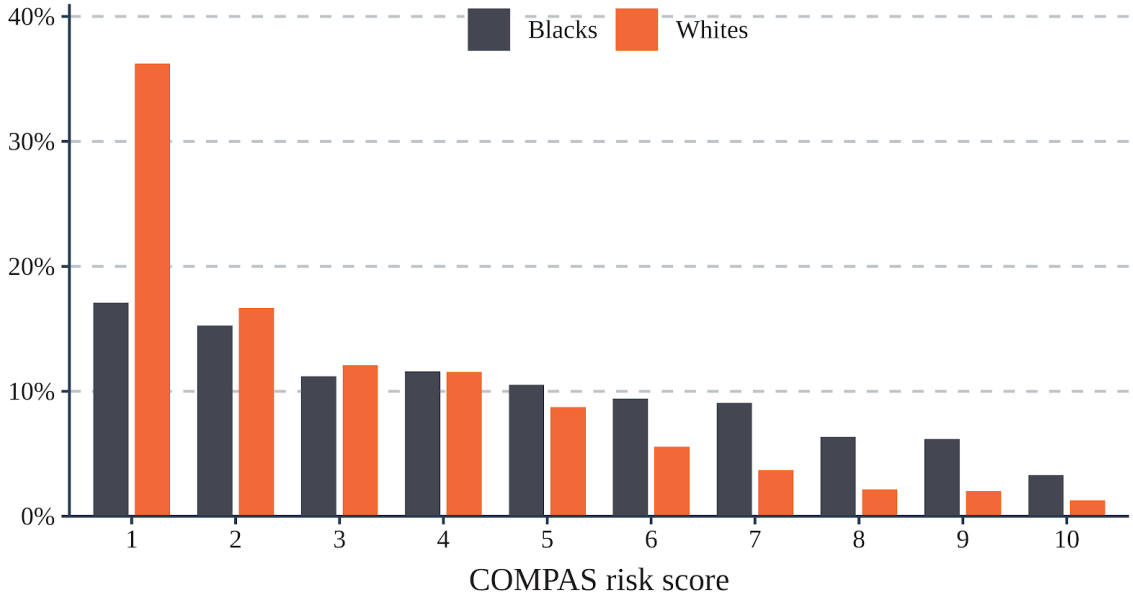

Figure 1: *The percentage of black and white defendants—who did not ultimately recidivate within two years—assigned to a COMPAS risk score.*

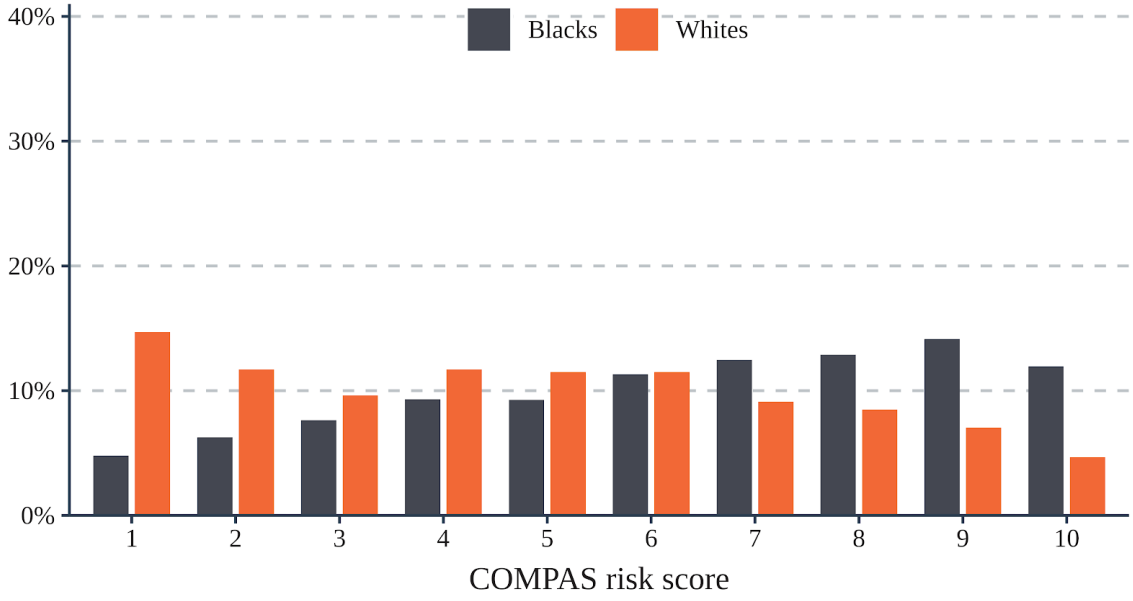

Figure 2: *The percentage of black and white defendants—who did ultimately recidivate within two years—assigned to each COMPAS risk score.*

A team from ProPublica—Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner—published what was at the time a groundbreaking piece of public journalism: using two years worth of ground truth data, they accused the algorithm of being biased against blacks, stating that of the defendants that *did not* ultimately recidivate, blacks were more than twice as likely to be classified as medium to high risk (risk score of 5–10) by the algorithm than whites. Conversely, of the defendants that *did* ultimately recidivate, whites were more than twice as likely to be classified as low risk (risk score of 1–4) by the algorithm than blacks (Angwin et al., 2016). This represents a breach of the *balance for the negative* class and the *balance for the positive class* respectively.

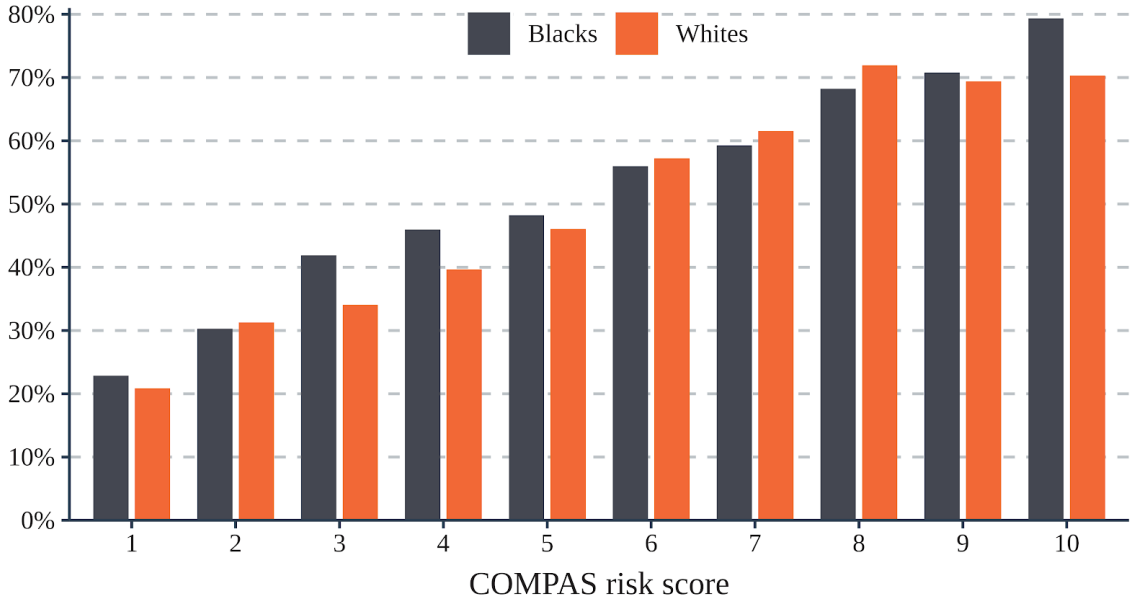

Figure 3: *The percentage of black and white defendants—assigned to each COMPAS risk score—that recidivated within two years.*

Northpointe responded arguing that within each bin, the algorithm was equally predictive for both blacks and whites (Dieterich et al., 2016). This represents a compliance of the *calibration within groups*.

Table 1: *The base rates of recidivism within two years for blacks and whites.*

Ethnicity	No. of Defendants	No. of Recidivists	Pct. of Recidivists (%)
Blacks	3,696	1,901	51.4
Whites	2,454	966	39.5

Figures 1, 2 and 3 illustrate the same data, sliced differently to evaluate Kleinberg et al.'s (2016) three notions of fairness: *balance for the negative class, balance for the positive class,* and *calibration within groups,* respectively. In the absence of *equal base rates* (see Table 1) and *perfect prediction,* it is straightforward to see how adjusting the algorithm to balance for one notion would disrupt the intricate balance of another.

How do we Tackle These Problems?

Upon first glance the solution seems straightforward: simply remove any information in the training dataset pertaining to the sensitive feature with respect to which we do not want the algorithm to be biased. However, it was found that the algorithms were able to probabilistically infer the sensitive feature using related information. Ergo, this approach saw little to no success.

One such case is Northpointe's COMPAS algorithm for predicting recidivism, discussed previously. Despite having removed any information related to race, the algorithm was able to infer race from other socioeconomic information such as income, educational background, type of residence, etc. Angwin et al. (2016) found the algorithm to be biased against blacks: of the defendants that did not ultimately reoffend, blacks were more than twice as likely to be classified as medium or high risk by the algorithm than whites.

Following this discovery, several methods for tackling algorithmic bias have been theorised and attempted. They can be broadly classified as either correcting the bias in the training process, or correcting the bias in the training dataset.

Correcting the Training Process

We previously pointed out that in machine learning, the core algorithm is hard to define. However, the designer can train the algorithm by adjusting the optimization criteria to include certain fairness criteria; but now we must enter turf that, once again, requires hard decisions and ethical quandaries.

In 2016, a team of computer scientists, with Moritz Hardt as the first author, explored some commonly used optimization criteria, and introduced two of their own: equal opportunity and equal odds.

- *Maximum profit.* This does not involve any fairness criteria and uses a different threshold for each group such that utility is maximised.
- *Race blind.* This uses a single threshold across all groups.
- *Demographic parity.* This uses a different threshold for each group such that the fraction of group members that are selected is the same across all groups.
- *Equal opportunity.* This uses a different threshold for each group such that the fraction of group members with a positive outcome that are selected is the same across all groups.
- *Equal odds.* This uses two different thresholds for each group such that the fraction of group members with a positive outcome that are selected is the same across all groups and the fraction of group members with a negative outcome that are selected is the same across all groups (Hardt et al., 2016).

Hardt et al. (2016) then compared these optimization criteria in the context of FICO credit scores using a TransUnion TransRisk dataset from 2003. Given that a threshold FICO score of 620 is commonly used for prime-rate loans and this corresponds to a loan default rate of 18%:

- The *maximum profit* model uses a different threshold FICO score for each group such that 82% of group members do not default. As one would expect, this model disadvantages blacks and hispanics.
- The race *blind model* uses a single threshold FICO score across all groups such that 82% of the population do not default. This model realises 99.3% of the profit available under the maximum profit model and still disadvantages blacks and hispanics.
- The *demographic parity* model uses a different threshold FICO score for each group such that the fraction of group members that qualify for a loan is the same across all groups. This model realises 69.8% of the profit available under the maximum profit model and results in reversing the bias such that whites and asians are disadvantaged.
- The *equal opportunity* model uses a different threshold FICO score for each group such that the fraction of non-defaulting group members that qualify for a loan is the same across all groups. This model realises 92.8% of the profit available under the maximum profit model.
- The *equal odds* model uses two different threshold FICO scores for each group that the fraction of non-defaulting group members that qualify for a loan and the fraction of defaulting group members that qualify for a loan is the same across all groups. Any group member above both thresholds qualifies for a loan, any group member below both thresholds does not qualify for a loan, and any group member in between both thresholds has a corresponding probability to qualify for a loan. This model realises 80.2% of the profit available under the maximum profit model.

Hardt et al. (2016) elaborated that the algorithm functions more accurately for majority groups than minority groups simply due to the abundance of training data. The equal opportunity model is able to utilise the algorithm's higher accuracy for majority groups, however the equal odds model constrains the algorithm to function as poorly as it does for minority groups for all groups.

Whilst this appears to be a promising framework, it is not without critique: it requires the selection of a type of bias, a trade-off that will have to be made. Liu et al. (2018) have gone further, arguing that when considering the system as a whole, of which the algorithm is simply a small part of, including fairness criteria in the optimisation criteria may result in harming the very groups it intends to protect by way of a delayed feedback loop. Once again, adopting the context of FICO credit scores, if a minority group member receives a loan after the introduction of fairness criteria that would not have been granted otherwise and ends up

defaulting on the loan, then his or her credit score would decrease, thus making it more difficult to acquire a loan in the future.

Correcting the Training Dataset

Correcting training data involves identifying all the different demographics and skews present in the training dataset and then adjusting the dataset to compensate. This approach has its proponents: researchers such as Chen et al. (2018) have argued that it is better to address the issue of bias by correcting the dataset rather than correcting the algorithm, as the latter involves a tradeoff in accuracy that is often unacceptable for sensitive applications such as healthcare. Feldman et al. (2015) claimed that bias prone datasets can be identified by building an algorithm that uses people's nonsensitive features to predict their sensitive features (ethnicity, sex, etc.). The accuracy with which these sensitive features can be predicted correlates to the extent to which an algorithm trained using the dataset can be biased; they proposed methods by which a dataset can be "repaired" so as to result in an unbiased algorithm yet retain relevant information.

Once again, these approaches involve ethically difficult choices to be made. Logically, to understand the bias against minorities and the underprivileged, certain protected classes of data—such as race and gender identity—must be examined; and for that, they must first be collected in a way that minimizes harm. Secondly, this comes back to the notions of what fairness is, something which will see some cases that are highly contextual—like credit scoring—and in some cases where contextual adaptation reduces critical accuracy. We believe it imperative that computer scientists work in interdisciplinary groups alongside social scientists for such efforts. We do not expect any easy answers.

Conclusion

While occurrences of algorithmic bias have been well documented, their origins traced, and their impacts acknowledged, much work remains to be done in unearthing and coming to consensus on potential solutions. The difficulty of this space extends outside merely establishing norms and ethics frameworks—every designer has to deal with theoretical and scientific limits as well as incongruous notions of what it means to be fair. Unless carefully considered, certain solutions to algorithmic bias may in actual fact exacerbate the issue. If ethical development is to meaningfully and profitably grapple with the space of algorithmic bias, these issues must be considered.

References

Knuth, D. E. (1968). *The art of computer programming.* Reading (Mass.) Menlo Park (Calif.) London etc: Addison-Wesley.

Markov, A. A. (1954). *Teoriya algerifmov* [Theory of algorithms]. Moscow: Academy of Sciences of the USSR.

Minsky, M. (1967). *Computation: finite and infinite machines.* Englewood Cliffs, N.J: Prentice-Hall.

Stone, H. S. (1972). Introduction to computer organization and structures. New York: McGraw-Hill.

Karp, R. M. (1986). Combinatorics, complexity, and randomness. Communications of the ACM, 29(2), 98–109. https://doi.org/10.1145/5657.5658

Thompson, A. (1996). An evolved circuit, intrinsic in silicon, entwined with physics. *Proceedings of the First International Conference on Evolvable Systems: From Biology to Hardware,* 390–405.

Karpathy, A. (2017, November 12). Software 2.0 [Blog post]. https://medium.com/@karpathy/software-2-0-a64152b37c35

Lee, N. T., Resnick, P., & Barton, G. (2019). *Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms.* https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). *Language Models are Unsupervised Multitask Learners.* https://openai.com/blog/better-language-models/

Wijeratne, Y. (2019, April 16). The Poetry Machine: Generating Tang Dynasty poetry using OpenAI GPT2 [Blog post]. https://towardsdatascience.com/the-poetry-machine-2764ec8b340b

Haskins, C. (2019, February 14). Academics Confirm Major Predictive Policing Algorithm is Fundamentally Flawed. *Vice.* https://www.vice.com/en_us/article/xwbag4/academics-confirm-major-predictive-policing-algorithm-is-fundamentally-flawed

Noble, S. (2018, March 26). Google Has a Striking History of Bias Against Black Girls. *Time.* https://time.com/5209144/google-search-engine-algorithm-bias-racism/

Hosanagar, K (2016, November 25). Blame the Echo Chamber on Facebook. But Blame Yourself, Too. *Wired.* https://www.wired.com/2016/11/facebook-echo-chamber/

Vincent, J. (2016, March 24). Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day. *The Verge*. https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist

Brown, M. (2016, October 6). Microsoft Japan's A.I. Teenage Chatbot Grows Depressed. *Inverse*. https://www.inverse.com/article/21827-microsoft-japan-rinna-ai-chatbot-yo-nimo-kimyo-na-monogatari

Google. (2017, August 25). *Machine Learning and Human Bias* [Video]. YouTube. https://www.youtube.com/watch?v=59bMh59JQDo

Cawley, G. C., & Talbot, N. L. C. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research, 11,* 2079–2107.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2019). A survey on bias and fairness in machine learning. *ArXiv:1908.09635 [Cs]*. http://arxiv.org/abs/1908.09635

Pasquale, F. (2015). *The Black box society: The secret algorithms that control money and information* (First Harvard University Press paperback edition). Harvard University Press.

Crawford, K. (2017, December 4-7). *The Trouble with Bias* [Keynote address]. Neural Information Processing Systems 2017. https://nips.cc/Conferences/2017/Schedule?showEvent=8742

Alciné, J. [@jackyalcine]. (2015, June 28). Google Photos, y'all fucked up. My friend's not a gorilla. [Tweet]. https://twitter.com/jackyalcine/status/615329515909156865

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. *ArXiv:1609.05807 [Cs, Stat]*. http://arxiv.org/abs/1609.05807

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May 23). Machine bias. *ProPublica*. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Dieterich, W., Mendoza, C., & Brennan, T. (2016). *COMPAS risk scales: Demonstrating accuracy equity and predictive parity.* https://www.equivant.com/response-to-propublica-demonstrating-accuracy-equity-and-predictive-parity/

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *ArXiv:1610.02413 [Cs].* http://arxiv.org/abs/1610.02413

Liu, L. T., Dean, S., Rolf, E., Simchowitz, M., & Hardt, M. (2018). Delayed impact of fair machine learning. *ArXiv:1803.04383 [Cs, Stat].* http://arxiv.org/abs/1803.04383

Chen, I., Johansson, F. D., & Sontag, D. (2018). Why is my classifier discriminatory? *ArXiv:1805.12002 [Cs, Stat].* http://arxiv.org/abs/1805.12002

Feldman, M., Friedler, S., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and removing disparate impact. *ArXiv:1412.3756 [Cs, Stat]*. http://arxiv.org/abs/1412.3756