Horizon Scanning Series

The Effective and Ethical Development of Artificial Intelligence: An Opportunity to Improve Our Wellbeing

Ethics, Bias and Statistical Models

This input paper was prepared by Oisín Deery and Katherine Bailey

Suggested Citation

Deery, O and Bailey, K (2018). Ethics, Bias and Statistical Models. Input paper for the Horizon Scanning Project "The Effective and Ethical Development of Artificial Intelligence: An Opportunity to Improve Our Wellbeing" on behalf of the Australian Council of Learned Academies, www.acola.org.

Ethical problems with statistical models

Dr. Oisín Deery, Lecturer, Department of Philosophy, Monash University
Katherine Bailey, Senior Manager, Artificial Intelligence, Accenture Technology

1 Introduction

In this article, we will briefly consider three questions and how they might be answered.

(a) Are there problems with using models based on population averages for predicting individual outcomes?
(b) How might using such models contribute to the further marginalization of already vulnerable population groups?
(c) What other sources of bias in statistical models should we be concerned with?

We will take these three questions in turn.

2 General problems

Let us address the first question:

(a) Are there problems with using models based on population averages for predicting individual outcomes?

Our answer to this question is "yes," with some caveats.

There are good and bad ways to use statistical models or algorithms to make predictions about individuals. For present purposes, a model is an abstract representation of a process, such as a bank's procedure for approving loan applications. We use the term "model" since the term "algorithm" picks out only the actual sequence of steps carried out during the training of a statistical model or at the time it issues predictions. Pedro Domingos, at the University of Washington, points out that in this sense of the term, algorithms cannot be biased. However, everyday use of the term "algorithm" generally picks out the trained model, which implicitly includes the data on which it was trained. If that data is biased, then there can be algorithms that embed biases, i.e., biased algorithms. Even so, for simplicity's sake we will use the term "model."

A statistical model takes data about the performance of variables as input, and assuming that this data is relevant to the outcomes the model is trying to predict, the model will make predictions about the outcomes based on the data. For example, a model might take data about factors judged as relevant to how likely people are to default on loans, and this model will then make predictions about how likely a particular individual is to default. Such a model may be used to rank loan applications, with those made by people predicted to have the highest likelihood of defaulting ranked lowest for approval, and those made by people predicted to have the lowest likelihood ranked highest (cf. Moritz et al. 2016).

Using a model in this way might seem to be a perfectly acceptable method for deciding whether to approve a loan, but only if certain conditions are met. First, the data on which the model relies must be suitably related to the performance of the variables that it is modeling. In the case of loan applications, the data should bear a direct relationship with whether people actually default on loans. It can be difficult to acquire such data, and in its absence modelers sometimes use what Cathy O'Neil

(2016) calls "proxy" data. For example, a model might rely on correlations between a person's postal code or language patterns and their likelihood of defaulting. Relying on correlations of this sort is discriminatory, and therefore wrong. Relatedly, another requirement is not satisfied in this case, namely that the data the model relies on be suitably related to whether people actually tend to default on loans.

Second, the data must be relevant to the outcomes the model is trying to predict. In cases where proxy data is used, this requirement is not satisfied either.

Third, a statistical model or algorithm should be able to learn from its mistakes, by having the capacity to self-correct in response to errors. Sometimes, errors in prediction can result from a model's relying on proxy data, while at other times errors can result from the model's making predictions from too little data. Thus, a model may predict that a particular person has a high likelihood of defaulting on a loan because this prediction was made on the basis of the person's postal code, even when the person in question has never defaulted on a loan and is therefore unlikely to do so now. Alternatively, a model might make a prediction about how likely a person is to do something based on too little data, in which case the prediction will be statistically unsound. Either way, without the ability to learn from feedback, a model will generate faulty predictions. Such a model will not only lack any way of learning whether it was correct in its predictions, but anyone who relies on the model may end up accepting its predictions as justification for using the model. In this way, a loan officer in a bank might justify continuing to use what is in fact a faulty model by pointing to the apparent fact that it has helped to prevent a bank's approving loans to people whom the model predicted would be likely to default, even if those predictions were in error.

Even when the criteria discussed above are satisfied, there can still be problems with using models based on population averages for predicting individual outcomes. For one thing, we need to remember that the predictions made by any model have a degree of uncertainty attached to them. Let us assume that someone's personal information is plugged into a model, and the model predicts the probability of this person's defaulting on a loan. If a high probability is returned, the person will be denied the loan. Even ignoring problems with how data is collected, and assuming perfect data about whether people tend to default on loans, models can sometimes make mistakes. This is because models are simplifications of whatever process is being modeled; no model can (or should) be complex enough to include information about all the variables that might actually bear on an actual-world process that we want to model. Models leave things out, and what they leave out will sometimes reveal the beliefs and goals of the people who designed it. So, for example, a model might deliberately be designed to sacrifice accuracy for efficiency. Because of this, the outputs of models often need to be taken with a grain of salt.

These issues have been considered in detail for various applications of statistical models or algorithms (see, e.g., Dwork et al. 2011; Feldman et al. 2014; Hardt et al. 2016).

3 Increasing marginalization

We will now address the second question:

(b) How might using such models contribute to the further marginalization of already vulnerable population groups?

As noted above, a model's relying on correlations between a proxy factor like a person's postal code and the likelihood of defaulting on a loan is discriminatory and ethically wrong. One reason it is wrong is that it further marginalizes people from already vulnerable groups. Most people living in a postal code where people are poor will already be vulnerable for various social and economic reasons, quite apart from their being additionally marginalized explicitly on the basis of the postal code in which they happen to live.

There are, however, other reasons why reliance on statistical models can exacerbate the marginalization of already vulnerable groups. For example, credit scores are sometimes used to rank applicants for a job, such that applicants with good credit scores are deemed preferable to those with bad scores. The justification for this practice might, for example, point to a correlation between people's having good credit scores and their tending to be punctual or follow rules, traits an employer might deem as desirable. The problem, however, is the uncertainty in making predictions on the basis of correlations. People who are punctual and follow rules can sometimes have bad credit scores merely because of bad luck. Yet these people will be penalized not only in being deemed relatively bad hires. They will be also penalized, in some cases, by the fact that they're not being hired will further decrease their credit score, in turn making it harder to gain employment. In this way, there can be a trade-off between a model's generating useful predictions and the way in which it might increase marginalization by inaccurately categorizing people sometimes.

When statistical models are used to inform hiring (and sometimes firing) decisions, further marginalization can occur. This is because statistical models are typically used in this way in the context of lower paid jobs, whereas hiring (and firing decisions) are more often made on the basis of personal judgment in the context of higher paid jobs. There are two problems here. First, those who have lower paid jobs will more often be at the mercy of the (often mistaken) outputs of such models, relative to those with higher paid jobs. Second, implicit biases in hiring practices for higher paid jobs will tend to prevent people from already vulnerable or marginalized groups from breaking into certain career pathways, since such biases will give people who already have higher paid jobs an advantage in such contexts.

The issues in relation to using statistical models in hiring practices, among other related use cases, have been considered by various researchers (see, e.g., Barocas et al. 2016; Hu & Chen 2017).

Bias in other contexts can be a problem too, sometimes even when statistical models are used with the laudable aim of reducing the influence of bias in human decision-makers. For example, statistical "recidivism models" aim at reducing the influence of bias (whether explicit or implicit) of judges in sentencing for crimes. But as various authors point out, these models in fact often serve only to reproduce and disguise bias. The data these models use is typically obtained from questionnaires that criminals are asked to complete. However, these data often include details about a criminal's upbringing, family, and social connections. Data of this sort ought to be irrelevant to a criminal's sentencing. Nevertheless, since the data is used to generate a "recidivism score" for criminals, and that score is used in sentencing, the data influences sentencing in a way that it should not. For example, a higher recidivism score will often result in a longer sentence. Even more perniciously, a longer sentence will often subsequently result in higher scores on the recidivism scale. Once again, we have a situation in which people from already marginalized groups in society can be further marginalized by statistical models.

The issues with using statistical models in relation to criminal sentencing have been dealt with in detail by Austin (2006), Vrieze & Grove (2010), Kort & Butters (2014), Angwin et al. (2016), Kehl et al. (2017), and Lum & Isaac (2016); compare Labrecque et al. (2014).

Relatedly, the healthcare industry now uses statistical techniques both in diagnosis and for identifying optimum treatments for patients. As two researchers recently noted in relation to these use-cases, "If undisclosed algorithmic decision-making starts to incorporate health data, the ability of black-box calculations to accentuate pre-existing biases in society could greatly increase" (Wilbanks & Topol 2016: 346). In the next section, we examine in more detail issue of biased data in healthcare applications.

4 Particular biases in statistical models

Let us now address our third question:

(c) What other sources of bias in statistical models should we be concerned with?

Here, we will consider three examples where bias in models can have negative effects in increasing marginalization of already marginalized and vulnerable groups in society.

4.1 Avoidable bias: dermatologist-level classification of skin cancer

The data used to train a model can be anything. In the case of healthcare diagnostic tools, the data will be patient information of various sorts. However, patients can be seriously harmed as a result of two features of how statistical models might be deployed as diagnostic tools in healthcare. First, there are inherent limitations in what statistical models can do, as well as various problems in how these models are trained, that make them artificially stupid. Second, there is widespread misunderstanding of what machine learning systems are capable of, with a resulting overconfidence in their outputs.

Consider the following headline: "Stanford's artificial intelligence is nearly as good as your dermatologist" (Mukherjee 2017). The story behind the headline was the striking success that researchers at Stanford University had in training a model to distinguish between benign, malignant, and non-neoplastic lesions in patients (Esteva et al. 2017). In other words, the Stanford researchers claimed to have developed a system that could detect whether a skin lesion is cancerous or not. But the artificial intelligence or "AI" system they developed is a statistical machine learning model—here, a model performing a supervised learning task, which is to classify images of lesions based on labeled images of lesions that it has previously seen. What these researchers did is remarkable. Yet it is simply not true that their system is anything like "nearly as good as your dermatologist."

First, a broad point. Artificial General Intelligence (aka AGI, or "strong AI") is often roughly defined as a computer system that is at least as intelligent as an average human being. The problem with headlines like this one is that they suggest we already have statistical models that have reached this benchmark. Yet even if Stanford's AI is nearly as good at diagnosing skin cancer as a dermatologist, the term "AI" is misleading here.

The Stanford system is good, at best, at only one thing: diagnosing skin cancer. It cannot do anything else. It is nowhere near being as flexibly intelligent as an average human being. The Stanford system is, in other words, a "weak AI," meaning that it is good at a specific task or range of tasks, as opposed to a "strong AI," which is what AI researchers call whatever future system (if any) might be as flexibly intelligent as an average human.

Even if we focus just on the specific task of identifying skin cancer, the problems we mentioned still arise. First, the Stanford model is only as good as the data that it is trained on. And second, that potential weakness of the system is too easily overlooked by both the medical profession and by healthcare policymakers, especially when the media suggests that people should defer to Stanford's system to (almost) the same extent as to a dermatologist. However, people should not defer to the model's outputs in this way. If hospital administrators or policymakers were to make the mistake of doing so, patients could easily be put at increased risk of harm, including death, as we will now illustrate.

The Stanford researchers trained their system by using photos from predominantly white people as training data. Their system was trained on images from three datasets, including the ISIC (or International Skin Imaging Collaboration) Archive. This dataset comprises 13,786 images of lesions that are biopsy-proven and labeled as malignant, benign, or non-neoplastic. In this dataset, there are only two or three images of lesions on black skin. Because the Stanford model was not trained on

images of lesions on black skin, it will simply not be able to reliably classify lesions in black patients. It could have learned to do this if it had been trained on appropriate data, but it was not.

It is doubtful the Stanford researchers were being explicitly racist in developing their model, although clearly, they were negligent (perhaps inadvertently) in not noticing or reporting this deficiency in their training data. Why did images of lesions on black skin not appear in the dataset?

It is unclear, but here are some hypotheses. First, fewer blacks than whites get skin cancer in the first place, partly as a result of natural UV protection in black skin (Glosser & Neal 2006). According to the Center for Disease Control and Prevention, for example, whites in the United States are almost thirty times more likely to present with skin cancer than blacks.[1] Thus, there are fewer cases of skin cancer among blacks to begin with. Worse, socioeconomic and medical insurance inequalities also contribute to fewer blacks presenting with lesions, mainly because they cannot afford to see a doctor, but also because of a relative lack of public education programs about skin cancer for blacks (Wich et al. 2011). Additionally, clinics at which blacks do present are typically the ones that tend to be less well-equipped to properly record lesions, since these clinics are often underfunded facilities in poorer neighborhoods. That such clinics are less well-equipped to adequately record lesions matters because the dataset on which the Stanford system was trained was developed partly with the aim of correcting for the fact that many medical images of lesions fail to meet minimum quality standards, including standards for adequate labeling. Thus, if existing images of lesions on black skin—many of which may have come from underfunded clinics—did not meet these standards, they would have been excluded from the dataset.

Whatever the explanation, the publicly available dataset on which the Stanford system was trained contains almost no images of lesions on black skin. By contrast, most dermatologists in the US receive extensive training in visually detecting skin cancer in blacks (despite—or perhaps because of—having fewer clinical cases to work with). As a result, human dermatologists can be expected to vastly outperform the Stanford system in visually detecting malignant lesions in blacks. Of course, this result did not show up during the testing of the Stanford system against twenty-one board-certified dermatologists, since images of lesions on black skin also did not feature in that test.

If policymakers at a hospital were to deploy a model of this sort to diagnose skin cancer, at least as an app that a patient can use on their smartphone before deciding to see a doctor, as has actually been proposed (Lofgreen et al. 2016), then a black patient might be given the all-clear, despite actually having skin cancer, since the model will be unable to reliably classify lesions on anything except white skin. At worst, the patient might die. Yet had the patient instead been examined by a qualified dermatologist (or even by a properly trained diagnostic model), he or she might have survived, by being diagnosed and treated earlier.

One might think that no dermatologist would be crazy enough to hand over the bulk of their diagnostic work to such a model. However, it has been considered. In a commentary published alongside the *Nature* paper reporting the Stanford model's results, Sancy Leachman, a researcher in the Department of Dermatology at Oregon Health and Science University, together with Glenn Merlino of the National Cancer Institute, responded by saying that systems like Stanford's could open up opportunities for doctors to use their time in ways that advance the field, instead of in diagnosis (Leachman & Merlino 2017). As Leachman said in an interview: "Let's spend our time and brain power on something that hasn't yet been solved," and leave what we already know to the computers. Thus, reliance on such models poses a genuine risk to patients' welfare.

To safeguard patients, statistical diagnostic models have to be better developed, better understood by the medical profession, and in most cases only deployed in conjunction with human medical expertise, in what researchers call the "human-in-the-loop" model (Holzinger 2016). In that case, models of this sort might indeed be useful.

[1] Source: https://www.cdc.gov/cancer/skin/statistics/race.htm

In summary, the dangers posed by models like the Stanford system have three sources. First, the medical profession (as well as the media) tends to overestimate what any plausible statistical model is capable of. Second, such overconfidence stems partly from people's tendency to think that statistical models are somehow inherently more objective than their human counterparts. This confidence is misplaced. Third, the confidence is also misplaced since the researchers who develop statistical models can make mistakes, one of the most dangerous of which is to allow bias to be introduced into the data on which the models are trained.

Despite the fact that the potential consequences of the Stanford model's being deployed are serious, the root problem is easily addressed: use better data. Nevertheless, it is worth reemphasizing that even when statistical models are trained on better data, their limitations should be better understood by the professionals who use them, especially when the models are proprietary or "black-box" in nature. In such models, the professionals who use them will not even know the basis for the outputs the system gives (Cohen et al. 2014). In the case of black-box models, unless a human is retained in the loop, deference to a model's output will differ little from deference to the proclamations of a supposedly divine oracle.

4.2 Difficult-to-avoid bias: predicting criminality

In other cases, it might not be clear how to train a model on better data. For example, researchers at Shanghai Jiao Tong University and McMaster University have claimed they have found evidence that criminality can be predicted from facial features. In their paper, Xiaolin Wu and Xi Zhang (2016) describe how they trained a model to be able to distinguish photos of criminals from photos of non-criminals with a high level of accuracy.

However, Wu and Zhang's results can be interpreted differently depending on what assumptions one brings to their paper, and what question one is interested in answering. The authors themselves simply assume, contrary to overwhelming evidence (e.g., Bobo & Thompson 2006), that there is no bias in the criminal justice system. Consequently, Wu and Zhang assume that the criminals whose photos they used as training data are a representative sample of the criminals in the wider population (including those who have never been caught or convicted for their crimes). The question in which Wu and Zhang are interested is whether there is a correlation between facial features and criminality. Given their assumption, they take their results as evidence that there is such a correlation.

Suppose, instead, that one starts from the assumption that there is no relationship between facial features and any putative criminality trait. In place of this question, one might instead be interested in whether there is bias in the criminal justice system. In that case, one will take Wu and Zhang's results as evidence that there is indeed such bias—i.e., that the criminal justice system is biased against people with certain facial features. After all, this hypothesis would also explain the difference between the photos of convicted criminals and the photos of people from the general population. The authors did not consider this alternative possibility. Indeed, they appear to be saying that while humans may be prone to bias, machine learning systems are not. They are explicit about this claim, in their response to critics:

> [S]ome of our critics seemed to suggest that machine learning tools cannot be used in social computing simply because no one can prevent the garbage of human biases from creeping in. We do not share their pessimism. Like most technologies, machine learning is neutral. (Wu & Zhang 2017)

However, it is clear that the data on which Wu and Zhang's system was trained had ample scope for human bias to enter at every step of the way, from the initial arrest to the conviction of each individual whose photograph appears in the dataset. The fact that Wu and Zhang dismiss this fact out of hand is disconcerting, to say the least. Indeed, they appear to suggest that we should deploy a system like this in the real world. Such a move could have disastrous consequences. For some purposes, such as

advertising, a false positive (an innocent person being identified as a criminal) would have few, if any, serious consequences. Yet for most purposes, a false positive could have ethically unacceptable results, e.g., unwarranted scrutiny of people who have done nothing wrong, or worse, arrests of innocent individuals.

Unlike in the case of the Stanford model for categorizing skin lesions, the bias in the data on which Wu and Zhang's model was trained seems difficult to eradicate, short of completely eradicating bias within the criminal justice system—a laudable aim that will probably not be achieved any time soon. Even then, it is unclear that the question Wu and Zhang set for themselves makes sense, since it is not clear what the trait of criminality is that their system is trying to measure, or how (even if there were such a trait) it might correlate with an individual's facial features.

Even more than in the case of Stanford's model for identifying skin cancer, one should resist any temptation one might feel to overestimate what a model like Wu and Zhang's is capable of. What Wu and Zhang fail to acknowledge is that human biases can infect the data on which supposedly "neutral" statistical models train, which will result in these models' also being biased. Indeed, were Wu and Zhang's model actually to be deployed in the world, it might even function perniciously to amplify the biases already present in the criminal justice system.

4.3 Impossible-to-avoid bias? NLP and word embeddings

In this section, we briefly consider yet another way in which bias might be introduced into the data on which statistical models are trained.

Despite great strides in natural language processing (NLP) by data-driven approaches, natural language understanding remains elusive. There is, however, a new technique for representing the words of a language that is proving incredibly useful in many NLP tasks, such as sentiment analysis and machine translation. The representations in question are known as "word embeddings," which involve mathematical representations of words that are trained from millions of examples of actual word usage. Word embeddings capture patterns of relationships between words in a multidimensional vector space. To use a classic example, a good set of representations would capture the relationship "king is to man as queen is to woman" by ensuring that a particular mathematical relationship holds between the respective vectors (specifically, king – man + woman = queen).

Such representations are at the heart of Google's new translation system, although they are representations of entire sentences, not just words. According to researchers at the Google Brain Team, this new system "reduces translation errors by more than 55%–85% on major language pairs measured on sampled sentences from Wikipedia and news websites" (Le & Schuster 2016) and can even perform zero-shot translations, i.e., translations between language pairs for which no training data exists.

However, researchers at Boston University and Microsoft Research (Bolukbasi et al. 2016) quickly noticed that Google's Word2Vec data set was apparently sexist. For example, just as the relationships "man is to woman as king is to queen," and "sister is to woman as brother is to man," were captured by word embeddings, so too were the relationships "man is to computer programmer as woman is to homemaker," and "father is to doctor as mother is to nurse."

The difficulty is that to work successfully, NLP systems relying on word embeddings need to learn the biases that exist in the bodies of text on which they are trained (Caliskan et al. 2017). After all, these biases are ours, and they are expressed in the instances of text on which the models in question are trained. Thus, if these models are successfully to learn the relationships that exist between words in our actual uses of language, they must learn relationships that are biased—including, for example, sexist relationships. Bias in the texts on which a model is trained are naturally going to be captured in the geometry of the word embeddings vector space. Worse, as Bolukbasi et al. put it, "The blind

application of machine learning runs the risk of amplifying biases present in data" (Bolukbasi et al. 2016).

Bolukbasi et al. describe a mathematical method, which they call "hard de-biasing." This method, they maintain, will partly reduce sexist and perhaps other unwanted bias (cf. Buolamwini & Gebru 2018) within the vector space, without compromising the overall structure—and thus the usefulness—of that space. However, they admit that their method is limited, and primarily works to reduce amplification of bias.

Of course, the best way to address the root of the problem might be obvious, and Bolukbasi et al. are aware of it. As they put it, "One perspective on bias in word embeddings is that it merely reflects bias in society, and therefore one should attempt to de-bias society rather than word embeddings" (Bolukbasi et al. 2016). However, that result is not something that can be achieved by means of a statistical model, if it can be achieved at all.

The important point is that even if Bolukbasi et al.'s method were to work for de-biasing word embeddings, some compensating measure (such as theirs) is needed to correct for the bias that these systems naturally acquire. So, vigilance is required. Both the developers and the users of any statistical model must resist any temptation they might feel to think that the model's outputs are more objective than the human-produced data on which it is trained. Additionally, developers and users must, in one way or another, take this fact into account, especially in cases where bias in the data seems impossible or difficult to eliminate.

References

Angwin, Julia, Jeff Larson, Surya Mattu & Lauren Kirchner. (2016). Machine bias. ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Austin, J. (2006). How much risk can we take? The misuse of risk assessment in corrections. *Federal Probation*, *70*(2): 58-63.

Bobo, Lawrence D. & Victor Thompson (2006). Unfair by design: The war on drugs, race, and the legitimacy of the criminal justice system. *Social Research: An International Quarterly*, *73*(2): 445–472.

Bolukbasi, Tolga, Kai-Wei Chang, James Zou, Venkatesh Saligrama & Adam Kalai (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. arXiv:1607.06520 [cs.CL]

Barocas, Solon & Andrew D. Selbst. (2016). Big data's disparate impact. *California Law Review*, *104*(671).

Buolamwini, Joy & Timnit Gebru. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of Machine Learning Research*, *81*: 1–15.

Caliskan, Aylin, Joanna J. Bryson & Arvind Narayanan (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334): 183–186.

Cohen, I.G., R. Amarasingham, A. Shah, B. Xie & B. Lo (2014). The legal and ethical concerns that arise from using complex predictive analytics in health care. *Health Affairs*, *33*(7): 1139–1147.

Dwork, Cynthia, Moritz Hardt, Toniann Pitassi, Omer Reingold & Rich Zemel. (2011). Fairness through awareness. arXiv:1104.3913 [cs.CC]

Esteva, Andre, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau & Sebastian Thrun (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, *542*: 115–118.

Feldman, Michael, Sorelle Friedler, John Moeller, Carlos Scheidegger & Suresh Venkatasubramanian. (2014). Certifying and removing disparate impact. arXiv:1412.3756 [stat.ML]

Glosser, Hugh M. & Kenneth Neal (2006). Skin cancer in skin of color. *Journal of the American Academy of Dermatology*, *55*(5): 741–760.

Hardt, Moritz, Eric Price & Nathan Srebro. (2016). Equality of opportunity in supervised learning. arXiv: 1610.02413 [cs.LG]

Hu, Lily & Yiling Chen. (2017). Fairness at equilibrium in the labor market." *Fairness, Accountability, and Transparency in Machine Learning*. arXiv:1707.01590 [cs.GT]

Kehl, Danielle, Priscilla Guo & Samuel Kessler. (2017). Algorithms in the criminal justice system: Assessing the use of risk assessments in sentencing. *Responsive Communities Initiative/ Berkman Klein Center for Internet & Society, Harvard Law School*. http://nrs.harvard.edu/urn-3:HUL.InstRepos:33746041

Labrecque, Ryan M., Paula Smith, Brian K. Lovins & Edward J. Latessa. (2014). The importance of reassessment: How changes in the LSI-R risk score can improve the prediction of recidivism. *Journal of Offender Rehabilitation*, *53*(2): 116–128.

Le, Quoc V. & Mike Schuster (2016). A neural network for machine translation, at production scale. *Google Research Blog*. https://research.googleblog.com/2016/09/a-neural-network-for-machine.htmachine learning

Leachman, Sancy & Glenn Merlino (2017). Medicine: The final frontier in cancer diagnosis. *Nature*, *542*: 119.

Lofgreen, Seth, Kurt Ashack, Kyle Burton & Robert Dellavalle (2016). Mobile device use in direct patient care. *Journal of the American Academy of Dermatology*, *74*(5): AB106.

Lum, Kristian & William Isaac. (2016). To predict and serve? *Royal Statistical Society*. doi: 10.1111/j.1740-9713.2016.00960.x

Mukherjee, Sy. (2017). Stanford's artificial intelligence is nearly as good as your dermatologist. *Fortune*. http://fortune.com/2017/01/26/stanford-ai-skin-cancer/

ONeil, Cathy. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Penguin Books.

Prince, Kort & Robert P. Butters. (2014). Brief report: An implementation evaluation of the LSI-R as a recidivism risk assessment tool in Utah. *The University of Utah/ Utah Criminal Justice Center*. https://socialwork.utah.edu/wp-content/uploads/sites/4/2016/11/LSI-R-Summary-Report-Final-v2.pdf

Vrieze, Scott I. & William M. Grove. (2010). Multidimensional assessment of criminal recidivism: Problems, pitfalls, and proposed solutions. *Psychological Assessment*, *22*(2): 382–395.

Wilbanks, John T. & Eric J. Topol. (2016). Stop the privatization of health data. *Nature*, *535*: 345–348.

Wich, Lindsay G., Michelle W. Ma, Leah S. Price, Stanislav Sidash, Russell S. Berman, Anna C. Pavlick, George Miller, Umut Sarpel, Judith D. Goldberg & Iman Osman (2011). Impact of socioeconomic status and sociodemographic factors on melanoma presentation among ethnic minorities. *Journal of Community Health*, *36*(3): 461–468.

Wu, Xiaolin & Xi Zhang (2016). Automated inference on criminality using face images. arXiv:1611.04135v3

Wu, Xiaolin & Xi Zhang (2017). Responses to critiques on machine learning of criminality perceptions. Addendum to arXiv:1611.04135