Patterns and Antipatterns, Principles, and Pitfalls: Accountability and Transparency in Artificial Intelligence

Jeanna Matthews

■ *This article discusses a set of principles for accountability and transparency in AI as well as a set of antipatterns or harmful trends too often seen in deployed systems. It provides concrete suggestions for what can be done to shift the balance away from these antipatterns and toward more positive ones.*

Increasingly, decisions that significantly impact the lives of individuals (such as decisions about hiring, housing, insurance, loans, criminal justice, or medical treatment) are being made in a partnership between human decision-makers and artificial intelligence (AI) systems. As builders of AI systems, we know how easy it is for errors to occur. We also know how difficult it can be to push the boundaries and adapt a system developed in one context into another. As developers of AI, we know how our systems learn from people and from the past, assimilating latent biases. Understanding all of this, who better than us to insist that the systems we build support investigation and iterative improvement, so that others are empowered to help counter the limitations of AI while benefiting from its strengths?

 ISSN 0738-4602

In 2017, the Association for Computing Machinery's US and European Public Policy Councils issued a statement of principles for algorithmic transparency and accountability (ACM Public Policy Council 2017; Garfinkel et al. 2017). Rather than a statement of specific ethics for systems,[1] it was focused on a set of seven principles to enable increased human oversight of systems. Enabling oversight allows us to build systems that can be inspected and compared with specific ethical standards.

This article begins by briefly surveying this set of principles. Next, I will describe a set of antipatterns, or harmful trends, seen too often when AI and machine learning systems are actually deployed. Finally, I will provide concrete suggestions for what can be done to shift the balance away from these antipatterns and toward more positive patterns.

Principles for Transparency and Accountability

The seven principles identified by Association for Computing Machinery (ACM)'s United States and European Public Policy Councils were awareness; access and redress; accountability; explanation; data provenance; auditability; and validation and testing. These principles are provided and explained in greater detail in the article by Garfinkel et al. (2017).

Awareness

Owners, designers, builders, users, and other stakeholders of analytic systems should be aware of the possible biases involved in their design, implementation, and use and the potential harm that biases can cause to individuals and society.

Access and Redress

Regulators should encourage the adoption of mechanisms that enable questioning and redress for individuals and groups that are adversely affected by algorithmically informed decisions.

Accountability

Institutions should be held responsible for decisions made by the algorithms that they use, even if it is not feasible to explain in detail how the algorithms produce their results.

Explanation

Systems and institutions that use algorithmic decision-making are encouraged to produce explanations regarding both the procedures followed by the algorithm and the specific decisions that are made.

Data Provenance

A description of the way in which the training data were collected should be maintained by the builders of the algorithms, accompanied by an exploration of the potential biases induced by the human or algorithmic data-gathering process.

Auditability

Models, algorithms, data, and decisions should be recorded so that they can be audited in cases where harm is suspected.

Validation and Testing

Institutions should use rigorous methods to validate their models and document those methods and results. In particular, they should routinely perform tests to assess and determine whether the model generates discriminatory harm.

Antipatterns

In this section, I consider what goes wrong when we fail to adhere to these principles. Specifically, I will consider a set of antipatterns that occur all too often in deployed systems.

Learning from the Past without Remembering the Context

Although we associate AI with future technology, it is interesting to consider the ways in which AI and machine-learning systems are really promoters and enforcers of the past. For example, we may train a system to recognize good candidates for a job by looking at data on who has successfully done that job in the past. Such a system might learn characteristics for successful computer programmers or nurses or chief executive officers that reflect the gender imbalances in those fields. This is possible even if the designers were to deliberately withhold columns such as race and gender from the input because there are many other possible proxies for these protected attributes.

For example, zip code can be used as proxy for race, magazine subscriptions for race or gender, purchasing patterns for medical conditions (Duhigg 2012; U.S. Consumer Financial Protection Bureau 2014; Allen 2018). Subtle differences in resumes such as mentioning softball versus baseball versus basketball versus polo versus sailing can signal difference in gender, race, and class even if less subtle clues like an applicant's name, gender, or address are withheld (Rivera and Tilcsik 2016). Amazon scrapped an internally developed recruiting engine when it downgraded graduates of women's colleges and resumes containing phrases such as "women's chess club captain" (Dastin 2018).

In his talk "Friends Don't Let Friends Deploy Black Box Models: Preventing Bias via Transparent Machine Learning," Rich Caruana warns against removing protected attributes prior to training. He warns that if offending bias variables are eliminated prior to training, then it both makes it harder tell when you still have a problem, and harder to correct the problems that remain. He recommends leaving bias features in data when the model is trained and then removing what was learned from these bias features after training. However, he notes that Article 9 of the General Data Protection Regulation covering the use of personal data revealing racial or ethnic origin and other

special categories might make this more difficult to do (Caruana 2017).

Beyond reproducing past human bias, AI systems can even amplify that bias. For example, Zhao et al. (2017) used a dataset of the images showing people cooking. In this dataset, the activity cooking was over 33 percent more likely to involve females than males, but a trained model further amplified the disparity to 68 percent at test time. Similarly, Douglas points out that using Google Translate to translate English text such as "he is a nurse. she is a doctor" into a language without gendered pronouns, such as Hungarian, and then back again to English, will produce text with the genders switched to "she is a nurse. he is a doctor." (Douglas 2017).

AI and machine-learning systems appropriately learn from the past, but the past is not a perfect oracle of the future that we want when data on the past reflects injustice and structural inequality. Users may view the decisions of deployed computer systems as fundamentally logical and unbiased, underestimating the degree to which the system may be encoding and even amplifying past human bias. It is important for human decision-makers to take this into account when considering the results produced by such systems.

Making Spurious Correlations

When machine learning systems look at data for patterns, it is easy for the system to identify attributes that may be correlated with the desired outcome, but that do not cause the outcome. A thought-provoking example of this comes from the article of Ribeiro et al. (2016), Why Should I Trust You? Explaining the Predictions of Any Classifier, in which they deliberately trained a classifier to differentiate between dogs and wolves by feeding it images of wolves surrounded by snow and dogs not surrounded by snow. When the system highlighted the portions of the images that were most influential in its decisions, it highlighted the snow as the reason for classifications. Without this addition, humans reviewing the classifications were much more likely to trust the classifications and much less likely to zero-in on snow as the spurious correlation. This illustrates the importance of techniques that provide explanations of recommendations for human review.

Learning from Humans without Remembering the Possibility of Malicious Training

Learning from past data are often learning indirectly from humans, for example, how humans have labeled instances in the past or what humans have considered successful in the past. It is also a common strategy to learn directly from humans, absorbing both the good and the bad of human behavior. Word embeddings trained on Google News articles exhibit substantial female and male gender stereotypes (Bolukbasi 2016). Microsoft's Tay, an AI chatbot released on Twitter in March 2016 is another example. Despite careful design, stress testing, and extensive user studies, Tay began producing what Microsoft Healthcare's vice president, Peter Lee, called "wildly inappropriate and reprehensible words and images" (Lee 2016). Lee speculated that this was the result of a coordinated attack, rather than Tay simply learning bad behavior from normal human usage patterns. Suciu et al. (2018) present a helpful overview of literature on poisoning machine learning in the context of a generalized model for the capabilities of adversarial agents.

Reusing Long Pipelines of Systems in Unanticipated Contexts

When developing a system, it is common to look for systems that can be used to reduce development burden. The developers of the original system may recognize limitations based on their design or training data or test coverage, but an appreciation of these limitations can easily be lost when a system is reused in a new and unanticipated context. Systems designed for one purpose have been reused in vastly different contexts such as a system designed for earthquake prediction used for predictive policing even though predictions do not influence the location of future earthquakes in the way the presence of police influences future arrest patterns (Goode 2011).

Employees at companies such as Google, Facebook, Amazon, and Microsoft have protested the proposed use of facial recognition systems in criminal justice or military applications (Eisikovits and Feldman 2018). While some employees focus on the ethics of those applications, others focus on the level of accuracy required for higher-stakes applications. Buolamwini and Gebru (2018) found that some commercial-grade facial recognition software had a 34.7 percent error rate for dark-skinned women, but only a 0.8 percent error rate for light-skinned men. O'Toole et al. found that algorithms developed in China, Japan, and South Korea recognized East Asian faces more accurately than Caucasian faces (O'Toole et al. 2011; Garvie and Frankle 2016) while Gyfcat, developed in Silicon Valley, reported problems with accuracy for Asian faces (Simonite 2018).

It is not surprising that the accuracy of facial recognition systems for different races is impacted by the demographics of the test cases used. However, given the wide disparity in accuracy, there is a good reason for concern when a system developed in one context is deployed in another context, especially in high-stakes contexts like criminal justice or military applications. Proposals like "Datasheets for Datasets" (Gebru et al. 2018) can help by providing a summary of when, where, and how the training data was gathered, its recommended use cases, and where applicable, information about the demographics and consent from human subjects.

As researchers and developers, we may build systems with the best intentions and using the best data we have available, but others cast about looking for a premade solution to plug into the empty hole in

the system they are building. We should do what we can to evangelize the limitations of our system, to prevent overzealous marketers from representing the system's features without a healthy respect for its limitations. Similarly, when we look for a premade solution to fit into the systems we are building, we should actively seek out and think critically about the possible limitations. For example, it is one thing to use a system that looks for warning signs of mental health issues to reach out to individuals with potentially helpful care, but it would be another thing to use that same system to deny employment or insurance coverage. In *Weapons of Math Destruction*: *How Big Data Increases Inequality and Threatens Democracy*, Cathy O'Neil describes the case of Ronald Behm who filed a class-action suit alleging the use of mental health screening questions during the job application process at companies like Finish Line, Home Depot, Kroeger, Lowe's, PetSmart, Walgreens, and Yum Brands after his son Kyle was unable to find employment and noted similarity of a common test used across employers to screening tests that had been used to diagnose his bipolar disorder (O'Neil 2016).

Using Inaccurate Data

Another big problem in deployed systems is the accuracy of the input data. Although not an AI-based system, the E-Verify system, designed to confirm the eligibility of employees to work in the U.S., is an example of the problem of decisions made based on erroneous input data. From 2006 to 2016, legal workers lost roughly 130,000 jobs and had their employment delayed for 580,000 more jobs due to E-Verify errors (Bier 2017). Many of these errors occurred because employers did not give individuals the opportunity to challenge or correct information in response to a tentative nonconfirmation decision (Westat 2012). In these examples, individuals have some ability to review decisions and characterizations made about them, but in many other cases, decisions are made without any review or opportunity to challenge the accuracy of the raw input data or even without individuals knowing that a decision is being made. One important level of explainability and transparency is to allow individuals to inspect and correct input data about themselves.

Using Data You Have Rather Than Data You Need or Missing Cases in the Data You Have

In addition to data used being inaccurate, the data available may simply not contain what you are trying to learn. It is tempting to use the data we have rather than insist on the data we actually need. O'Neil describes the tendency to use credit score as a proxy for being a responsible person, without an attempt to establish cause or even correlation (O'Neil 2016). She points out that there are many reasons responsible, trust-worthy employees may have a poor credit score and using credit scores in this way, simply because they are readily available, sets up a self-enforcing cycle of poverty (for example, if you can't get a job, your credit score will get worse).

More generally, Ben Green points out that machine learning systems grant undue weight to quantified considerations at the expense of unquantified ones (Green 2018). In the context of recidivism risk assessment tools in the criminal justice system, judges may place greater emphasis on incapacitating offenders from committing further crimes rather than on important goals of sentencing such as deterring others from committing similar crimes in the future, rehabilitating offenders, and delivering just punishment. He observes that deployment of a tool or algorithm to quantify one aspect may distort the values underlying laws and policies without review or proper democratic input.

Allowing individuals to present additional or alternate evidence that may more directly support the conclusion desired would be an excellent step. This would require human decision-makers able to weigh the alternate evidence in conjunction with those features the computer system is designed to consider.

In another unfortunately common variant of this antipattern, systems can fail in exceptional cases that are simply not covered by the training data. Without proper explanation and review, these problems can go undiagnosed especially if systems designers are satisfied with aggregate system performance and unwilling to invest resources in debugging rare or individual cases in which the system delivers bad results. Enabling impacted individuals to request investigation and incentivizing human decision-makers to investigate reports of bugs or unexpected outcomes in individual cases would help with this problem. Roselli, Matthews, and Talagala (2019) discuss ways to manage bias including instrumenting systems to highlight when production data are substantially different than the data seen in training.

Aggressively Resisting Review

Many of the problems I have discussed are exacerbated by a tendency of software vendors to aggressively resist review including with Dewitt clauses in Terms of Service that attempt to prevent the publication of benchmark results without the permission of the manufacturer, or laws such as the Computer Fraud and Abuse Act and the Digital Millennium Copyright Act that software vendors have used to bring legal action against researchers for allegedly exposing defects in their systems (Doctorow 2018). These laws have a real and chilling impact on research that is needed to reveal problems and provide incentives to improve (Felten 2013; Wilson and Mislove 2017).

Even defense experts in criminal cases are regularly denied access to source code and system details under protective order when software vendors claim trade secret protection (Wexler 2018). For example, in New York City, defense experts were for years denied access to the source code of Forensic Statistical Tool (FST),[3] a DNA genotyping software system used to match evidence samples to suspect's DNA. When

permission for the first source code level review was granted under protective order, the findings of trouble with software quality and undisclosed methods that discarded data of possible use to the defense were stunning enough that the expert's findings were released in fully unredacted form and the source code was even released publicly and posted on GitHub by Propublica (Kirchner 2017; Matthews et al. 2019). Similarly, in Idaho, requests to explain drops in Medicaid benefits were initially met with claims of trade secret protection and then later forced disclosures revealed incorrect input data, undisclosed troubles with internal testing, and fundamental statistical flaws in the underlying formula used (Stanley 2017). Intellectual property protection should be used as a reward for great ideas, not as a way to avoid embarrassing results that would legitimately point to bugs, bias, and other problems in a system.

Without the opportunity to conduct third-party investigation and review, software developers are often not sufficiently incentivized to disclose possible problems and invest in improvement of systems. Validation and testing may be done with the aim of documenting the success of the system rather than rooting out corner-case errors that can have substantial negative impacts for individuals. When we see a bug in a computer system, even if that bug does not seem to impact others around us, we expect that eventually the bug will be identified and fixed. However, when the organizations purchasing the software have different interests than the individuals about whom decisions are being made, market forces alone may be insufficient to incentivize this iterative improvement. For example, if there is an error in criminal justice software, would a bug report be truly investigated or would the response be that you're just complaining because you are guilty? Without adversarial testing and third-party review, what hope would there be of a real error being found and fixed?

Failing to Measure the Social Impact of Deployed Systems

No matter how thoroughly designed and tested, systems are bound to cause damage if they fail to measure the impact of deployed systems. Virginia Eubanks offers a number of powerful examples of this in her book *Automating Inequality*: *How High-Tech Tools Profile*, *Police and Punish the Poor*, including the automation of Indiana's welfare eligibility system and the development of the Allegheny Family Screening Tool to predict child abuse and neglect (Eubanks 2017). In Indiana, many people lost critical services when errors in the new system made it almost impossible for them to submit required paperwork even though any deviation from the newly rigid application process was interpreted as an active refusal to cooperate and resulted in cancellation of benefits.

Systems are typically implemented to achieve cost-savings, efficiency, and reduced risks for decision-makers. However, the interests of society as a whole can be very different. For example, in many high stakes areas such as hiring, criminal justice, and the allocation of public resources, there are legal and moral obligations that require the effort to review individual cases carefully. The societal framework for these obligations has been worked out over a long period of time, but in the process of automating certain decisions, we could shift some of our fundamental societal values without discussion or review. Green discusses this in the context of risk assessment tools used in the criminal justice system (Green 2018), but it is applicable in many other areas. We may think we are simply automating an existing process for efficiency or cost-savings, but developers may make a wide variety of implementation decisions that are no longer highlighted to human decision-makers or to society as a whole, as options. It isn't easy to have it all, both cheap, efficient decision-making and careful, custom individual consideration. It is important to grapple with what is being lost in the process of automation and build-in robust review processes to focus human decision-makers on cases that need more customized consideration.

As AI researchers, we should worry not just about increased efficiency in decision-making for those purchasing systems, but also about individuals impacted by the system and about its impact on society as a whole. We can look to human rights law, documents of professional ethics like the ACM Code of Ethics and Professional Conduct (ACM 2018; Gotterbarn 2017), and statements such as The Toronto Declaration: Protecting the Rights to Equality and Non-Discrimination in Machine Learning Systems (Amnesty International and Access NOW 2018) for some good inspiration. Selbst and Powles (2017) discuss the requirements for explanation ("meaningful information about the logic involved" in automated decisions) under the General Data Protection Regulation and Latonero (2018) discusses governance of AI from the perspective of human rights and dignity.

Inappropriately Defining the Role and Responsibility of Humans in the Decision-Making Process

Big decisions about our lives are increasingly made jointly by humans and computer systems, but what role should human decision-makers play? Too often, computer systems are used to relieve human decision-makers of the moral cost of making difficult choices, like whom to fire or whom to send to prison. Human decision-makers are encouraged to conclude "I just do what the computer tells me," allowing the system to absorb blame. Even in high agency professions like doctors or judges, it is important to question what freedom humans will have to overrule the computer without risking, for example, malpractice suits or a record that looks soft on crime.

Alternatively, Elish (2019) uses the term *moral crumple zone* to describe the opposite situation when humans in the loop can be used to absorb legal and moral liability when automated systems fail. She describes how human pilots were blamed after they

failed to recover from a stall when the autopilot system shut itself off, causing the fatal crash of Air France Flight 447 and argues that an increase in automation can make pilots' skills atrophy.

Human oversight may be better deployed investigating reports of errors than rubber-stamping each decision. Automated systems can reduce labor costs related to decision-making, but some of that savings could be reallocated to the investigation of problems. Teams of human decision-makers, capable of changing the outcome, could be tasked with investigating reports of problems and truly rewarded when bugs are identified. Systems could be instrumented to provide human investigators with the detailed information necessary to serve this important role.

It is inspiring to imagine that we could craft workflows that know when to draw on human intelligence and when to draw on computer intelligence, so that we can benefit from the best of both worlds. However, without an investment in improving the explainability of AI, we may end up with the worst of both.

Providing Transparency without Specifying Accountability

It is important for researchers and systems builders to invest in tools that open up the black box and help them identify flaws and critically assess unintended consequences of their systems. However, beyond information, we need controls that allow the system to be more accountable to stake holders and society as a whole. An understanding of the flaws must be coupled with the will and processes to improve it. In their article Accountable Algorithms, Kroll et al. (2017) discuss the governance of algorithms and ways to ensure the interests of citizens, and society as a whole.

The more important the decision, the more important it is to provide an explanation for it; high stakes decisions, especially in regulated areas like hiring and housing or in public policy contexts like criminal justice or the allocation of public resources, deserve more explanation. There are often tradeoffs between explainability and accuracy, but we have to consider not just the risk of lower accuracy but also the risk that a black-box system contains errors or even malicious content that could be exposed through an investment in explainability. For high stakes decisions, explanation may be even more important than an improvement in accuracy from a less explainable algorithm. This is especially true for complex systems in which portions are developed elsewhere and reused in a new context. Organizations and individuals using automated systems as a tool need enough information that they can explain, and ultimately be held accountable for the decisions they are making.

What Can We Do?

I have reviewed a set of seven desirable principles for algorithmic accountability and transparency as well as 10 antipatterns seen in deployed systems. This section is a wish list of actions that we as scientists and researchers could take to enable transparency and accountability and to create incentives for incremental improvement of systems rather than black boxes.[2] This list challenges all of us to ask if we are doing everything we need to do as responsible scientists to clearly demonstrate the weaknesses in our systems, to encourage users of our systems to retain an appropriate skepticism of the results, and to enable the people impacted by our systems to challenge them.

First, we should actively highlight the assumptions and limitations to users of our systems and to other developers who consider using our system as a building block.

Second, we should use documents like the ACM Code of Ethics and Professional Conduct, the USACM/EU-ACM Statement on Accountability and Transparency, and The Toronto Declaration to lobby for changes within our organizations and to justify the need to invest in instrumentation of our systems to explain their outputs to all stakeholders impacted by the system (ACM US Public Policy Council 2017; Garfinkel et al. 2017; ACM 2018; Amnesty International and Access NOW 2018; Gotterbarn et al. 2017).

Third, we should prioritize research into new ways to provide explanations and transparency, especially for currently deployed systems that are not currently amenable to explanation. New techniques for explanation and transparency are an active and encouraging area of research (Datta, Sen, and Zick 2016; Ferreira, Zafar, and Gummadi 2016; Lei, Barzilay, and Jaakkola 2016; Ribeiro, Singh, and Guestrin 2016; Pei et al, 2017; Dhurandhar et al. 2018).

Fourth, we should distinguish clearly between learning from the past and reproducing it, and remind those who use systems trained on past data to consider more than what has worked in the past when making decisions about the future.

Fifth, we should actively ask what the mechanisms and incentives for identifying and correcting flaws in our systems will be, once deployed. We should establish teams of humans to receive and investigate reports of errors and reward them when errors are found. We should encourage external third-party testing and enable automated third-party testing with clear, scriptable interfaces. Barriers to public scrutiny should be avoided whenever possible, especially in regulated areas.

Sixth, and finally, we should provide a way for individuals negatively impacted by our systems to seek effective recourse, including inspecting input data about them and providing additional evidence for consideration. We should actively consider our legal and moral obligations not just to purchasers of our systems, but also to those about whom decisions are being made.

Summary

As researchers, we love to envision the real-world scenarios in which our research could offer recipes

for improving the world. However, the truth is that powerful technologies are rarely used for good only, and AI is no exception. Although envisioning the ways in which our work could unintentionally lead to harm is not as enjoyable, this article has presented several ideas and suggestions for promoting a culture of AI research in which researchers can play as active a role in controlling the potential misuse of AI as they do in advancing its potential for good.

Acknowledgments

I thank my colleagues in the ACM's Technology Policy Committees for their many contributions to the Statement on Algorithmic Transparency and Accountability as well as to the formation of these ideas overall. I also thank my colleagues at Data and Society, where I was a 2017–2018 Fellow.

Notes

1. Discussion of ethical principles for AI systems is essential, but it brings subtle consideration of the relative importance of the well-being of individual humans versus collective societal well-being or even the well-being of the earth and other living things. Different cultures and societies around the world vary in how they prioritize these and other ethical values when they come into conflict.

2. Different actors (researchers, system developers, system deployers, policy makers, and citizens) have different abilities to influence action in this space. In this article, I am encouraging people in these roles (and others) to consider what they can, but not specifically exploring the limitations and possibilities of what action is possible in each role.

3. www1.nyc.gov/assets/ocme/downloads/pdf/technical-manuals/protocols-for-forensic-str-analysis/forensic-statistical-tool-fst.pdf.

References

ACM. 2018. *ACM Code of Ethics and Professional Conduct*. New York: Association for Computing Machinery. ethics.acm.org/2018-code-draft-3/

ACM US Public Policy Council. 2017. *Statement on Algorithmic Transparency and Accountability*. New York: Association for Computing Machinery. www.acm.org/binaries/content/assets/public-policy/2017_usacm_statement_algorithms.pdf

Allen, M. 2018. Why Health Insurers Track When You Buy Plus-Size Clothes or Binge-Watch T. *PBS News Hour,* July 17. www.pbs.org/newshour/health/why-health-insurers-track-when-you-buy-plus-size-clothes-or-binge-watch-tv

Amnesty International and Access NOW. 2018. The Toronto Declaration: Protecting the Rights to Equality and Non-Discrimination in Machine Learning Systems. May 16. www.accessnow.org/the-toronto-declaration-protecting-the-rights-to-equality-and-non-discrimination-in-machine-learning-systems/

Bier, D. 2017. E-Verify Has Delayed or Cost Half a Million Jobs for Legal Workers. May 16. Washington, DC: Cato Institute. www.cato.org/blog/e-verify-has-held-or-cost-jobs-half-million-legal-workers

Bolukbasi, T.; Chang, K.; Zou, J.; Saligrama, V.; and Kalai, A. 2106. Man Is to Computer Programmer as Woman Is to Homemaker? Debiasing Word Embeddings. Paper presented at the 30th International Conference on Neural Information Processing Systems, December 5–10, Barcelona, Spain.

Buolamwini, J., and Gebru, T. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. Proceedings of the 1st Conference on Fairness, Accountability and Transparency. *Proceedings of Machine Learning Research* 81(1): 77–91. proceedings.mlr.press/v81/buolamwini18a.html.

Caruana, R. 2017. Friends Don't Let Friends Deploy Black Box Models: Preventing Bias via Transparent Machine Learning. Invited Talk presented at the 2017 Fairness, Accountability and Transparency in Machine Learning (FATML) Conference, August 14, Halifax, NS. www.fatml.org/media/documents/2017_rich_caruana_friends_dont_let_friends_deploy_blackbox_models.pdf.

Dastin, J. 2018. Amazon Scraps Secret AI Recruiting Tool That Showed Bias Against Women. Reuters Business News, October 10, 5:00 AM. www.reuters.com/article/amazoncom-jobs-automation/rpt-insight-amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSL2N1WP1RO.

Datta, A.; Sen, S.; and Zick, Y. 2016. Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems. In *Proceedings of 2016 IEEE Symposium on Security and Privacy*. Piscataway, NJ: Institute for Electrical and Electronics Engineers. doi.org/10.1109/SP.2016.42.

Dhurandhar, A.; Chen, P.; Luss, R.; Tu, C.; Ting, P.; Shanmugam, K.; and Das, P. 2018. *Explanations Based on the Missing: Towards Contrastive Explanations with Pertinent Negatives*. arXiv:1802.07623. [cs.AI]. Ithaca, NY: Cornell University Library.

Doctorow, C. 2018. Telling the Truth About Defect in Technology Should Never, Ever, Ever Be Illegal, EVER. Electronic Frontier Foundation, August 15. www.eff.org/deeplinks/2018/08/telling-truth-about-defects-technology-should-never-ever-ever-be-illegal-ever.

Douglas, L. 2017. AI Is Not Just Learning Our Biases; It Is Amplifying Them. Medium, December 5. medium.com/@laurahelendouglas/ai-is-not-just-learning-our-biases-it-is-amplifying-them-4d0dee75931d.

Duhigg, C. 2012. How Companies Learn Your Secrets. New York Times Magazine. February 16. www.nytimes.com/2012/02/19/magazine/shopping-habits.html.

Eisikovits, N., and Feldman, D. 2018. Employees at Google, Amazon and Microsoft Have Threatened to Walk Off the Job Over the Use of AI. The National Interest Buzz Blog, August 5. nationalinterest.org/blog/buzz/employees-google-amazon-and-microsoft-have-threatened-walk-job-over-use-ai-27962.

Elish, M. 2019. Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction. *Engaging Science. Technology in Society* 5: estsjournal.org/index.php/ests/article/view/260.

Eubanks, V. 2017. *Automating Inequality: How High-Tech Tools Profile, Police and Punish the Poor*. New York: St. Martin's Press.

Felten, E. 2013. The Chilling Effects of the DMCA. Slate. March 29. www.slate.com/articles/technology/future_tense/2013/03/dmca_chilling_effects_how_copyright_law_hurts_security_research.html.

Ferreira, M.; Zafar, M.; and Gummadi, K. 2016. *The Case for Temporal Transparency: Detecting Policy Change Events in Black-Box Decision Making Systems*. arXiv preprint. arXiv:1610.10064. [cs.AI]. Ithaca, NY: Cornell University Library.

Garfinkel, S.; Matthews, J.; Shapiro, S.; and Smith, J. 2017. Toward Algorithmic Transparency and Accountability. *Communications of the ACM* 60(9): 5. doi.org/10.1145/3125780.

Garvie, C., and Frankle, J. 2016. Facial-Recognition Software Might Have a Racial Bias Problem. *Atlantic (Boston, Mass.)* 7(April): www.theatlantic.com/technology/archive/2016/04/the-underlying-bias-of-facial-recognition-systems/476991/.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Wortman Vaughan, J.; Wallach, H.; Daumeé, H.; and Crawford, K. 2018. *Datasheets for Datasets*. arXiv preprint. arXiv:1803.09010. Ithaca, NY: Cornell University Library.

Goode, E. 2011. Sending the Police Before There's a Crime. New York Times, August 15. www.nytimes.com/2011/08/16/us/16police.html.

Gotterbarn, D.; Bruckman, A.; Flick, C.; Miller, K.; and Wolf, M. 2017. ACM Code of Ethics: A Guide For Positive Action. *Communications of the ACM* 61(1): 121–8. doi.org/10.1145/3173016.

Green, B. 2018. Fair Risk Assessments: A Precarious Approach for Criminal Justice Reform. Paper presented at the Fairness, Accountability, and Transparency in Machine Learning (FATML) 2018 conference, July 15, Stockholm, Sweden. scholar.harvard.edu/bgreen/publications/%E2%80%9Cfair%E2%80%9D-risk-assessments-precarious-approach-criminal-justice-reform.

Kirchner, L. 2017. Federal Judge Unseals New York Crime Lab's Software for Analyzing DNA Evidence. Propublica, October 20. www.propublica.org/article/federal-judge-unseals-new-york-crime-labs-software-for-analyzing-dna-evidence.

Kroll, J.; Huey, J.; Barocas, S.; Felten, E.; Reidenberg, J.; Robinson, D.; and Yu, H. 2017. Accountable Algorithms. *University of Pennsylvania Law Review* 165(3): 633–706.

Latonero, M. 2018. *Governing Artificial Intelligence: Upholding Human Rights and Dignity*. New York: Data and Society Institute, datasociety.net/wp-content/uploads/2018/10/DataSociety_Governing_Artificial_Intelligence_Upholding_Human_Rights.pdf.

Lee, P. 2016. Learning from Tay's Introduction. *Official Microsoft Blog,* March 25. blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/#sm.00001fz4tifqfod9iszo7nfwvhi8s.

Lei, T.; Barzilay, R.; and Jaakkola, T. 2016. *Rationalizing Neural Predictions*. arXiv preprint. arXiv:1606.04155. [cs.CL]. Ithaca, NY: Cornell University Library.

Matthews, J.; Lorenz, S.; Babaeianjelodar, M.; Matthews, A.; Njie, M.; Adams, N.; Krane, D.; Goldthwaite, J.; and Hughes, C. 2019. The Right to Confront Your Accusers: Opening the Black Box of Forensic DNA Software. In *Proceedings of the 2019 AAAI/ACM Conference on Artificial Intelligence, Ethics and Society (AIES)*, 321–7. New York: Association for Computing Machinery. doi.org/10.1145/3306618.3314279.

O'Neil, C. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York: Random House.

O'Toole, A.; Phillips, P.; An, X.; and Dunlop, J. 2011. *Demographic Effects on Estimates of Automatic Face Recognition Performance. NISTIR 7757*. Gaithersburg, MD: U.S. Dept. of Commerce, National Institute of Standards and Technology doi.org/10.6028/NIST.IR.7757.

Pei, K.; Cao, Y.; Yang, J.; and Jana, S. 2017. DeepXplore: Automated Whitebox Testing of Deep Learning Systems. In *Proceedings of the 26th Symposium on Operating Systems Principles (SOSP 17)*, 1–18. New York: Association for Computing Machinery. doi.org/10.1145/3132747.3132785.

Ribeiro, M.; Singh, S.; and Guestrin, C. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–44. New York: Association for Computing Machinery. doi.org/10.1145/2939672.2939778.

Rivera, L., and Tilcsik, A. 2016. Class Advantage, Commitment Penalty: The Gendered Effect of Social Class Signals in an Elite Labor Market. *American Sociological Review* 81(6): 1097–131. journals.sagepub.com/doi/abs/10.1177/0003122416668154.

Roselli, D.; Matthews, J.; and Talagala, N. 2019. Managing Bias in AI. In *Companion Proceedings of the 2019 World Wide Web Conference*, 539–544. New York, NY: Association for Computing Machinery.

Selbst, A., and Powles, J. 2017. Meaningful Information and the Right to Explanation. *International Data Privacy Law* 7(4): 233–42. doi.org/10.1093/idpl/ipx022.

Simonite, T. 2018. How Coders Are Fighting Bias in Facial Recognition Software. Wired Business, March 29. www.wired.com/story/how-coders-are-fighting-bias-in-facial-recognition-software.

Stanley, J. 2017. *Pitfalls of Artificial Intelligence Decisionmaking Highlighted In Idaho ACLU Case. Free Future. American Civil Liberties Union,* June 2. www.aclu.org/blog/privacy-technology/pitfalls-artificial-intelligence-decisionmaking-highlighted-idaho-aclu-case.

Suciu, O.; Marginean, R.; Kaya, Y.; Daume, H.; and Dumitras, T. 2018. When Does Machine Learning {FAIL}? Generalized Transferability for Evasion and Poisoning Attacks. Paper presented at the 27th USENIX Security Symposium, August 15–17, Baltimore, MD. www.usenix.org/node/217487.

U.S. Consumer Financial Protection Bureau. 2014. Using Publicly Available Information to Proxy for Unidentified Race and Ethnicity: A Methodology and Assessment. *Industry and Markets Report,* September 17.files.consumerfinance.gov/f/201409_cfpb_report_proxy-methodology.pdf.

Westat Corporation. 2012. *Evaluation of the Accuracy of E-Verify Findings*. Rockville, MD: Westat. www.e-verify.gov/sites/default/files/everify/data/FindingsEVerifyAccuracyEval2012.pdf.

Wexler, R. 2018. Life, Liberty, and Trade Secrets, Intellectual Property in the Criminal Justice System. *Stanford Law Review* 70(4): 1343–429.

Wilson, C., and Mislove, A. 2017. We're Suing the Federal Government to Be Free to Do Our Research. The Conversation, March 27. theconversation.com/were-suing-the-federal-government-to-be-free-to-do-our-research-74676.

Zhao, J.; Wang, T.; Yatskar, M.; Ordonez, V.; and Chang, K. 2017. Men Also Like Shopping: Reducing Gender Bias Amplification Using Corpus-Level Constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Stroudsberg, PA: Association for Computational Linguistics. doi.org/10.18653/v1/D17-1323.

Jeanna Neefe Matthews is an associate professor of computer science at Clarkson University. She is a member of the ACM's Technology Policy Council and the chair of the AI and Algorithmic Accountability subcommittee. She is an affiliate at Data and Society in Manhattan.