Managing Bias in AI

Drew Roselli
ParallelM
Sunnyvale, CA, USA,
drew.roselli@parallelm.com

Jeanna Matthews
Department of Computer Science
Clarkson University, Potsdam, NY,
USA, jnm@clarkson.edu

Nisha Talagala
Saratoga
CA, nishatalagala@gmail.com

ABSTRACT

Recent awareness of the impacts of bias in AI algorithms raises the risk for companies to deploy such algorithms, especially because the algorithms may not be explainable in the same way that non-AI algorithms are. Even with careful review of the algorithms and data sets, it may not be possible to delete all unwanted bias, particularly because AI systems learn from historical data, which encodes historical biases. In this paper, we propose a set of processes that companies can use to mitigate and manage three general classes of bias: those related to mapping the business intent into the AI implementation, those that arise due to the distribution of samples used for training, and those that are present in individual input samples. While there may be no simple or complete solution to this issue, best practices can be used to reduce the effects of bias on algorithmic outcomes.

CCS CONCEPTS

• Computing methodologies~Artificial intelligence • Computing methodologies~Machine learning • Security and privacy~Social aspects of security and privacy

KEYWORDS

Artificial intelligence; bias; production monitoring

ACM Reference format:

Drew Roselli, Jeanna Matthews and Nisha Talagala. 2019. Managing Bias in AI. In *Proceedings of WWW '19: The Web Conference (WWW '19), May 13, 2019, San Francisco, USA*. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3308560.3317590

1. Introduction

The potential for efficiencies and improvements has attracted many companies to invest in AI [1] [2]. However, the specter of unintentional bias and its potential to cause harm can negatively impact a company's reputation. This type of risk and consequent exposure to lawsuits is an important reason for caution. Bias in algorithms not only carries these risks, it can cause an application to perform sub-optimally, leading to missed opportunities. For example, bias in lending practices can lead to both financial loss, when bias favors some people, and missed financial gain when bias unfairly discriminates against others. The situation is exacerbated by the fact that many AI systems are not "explainable" in ways that deployers can claim to not contain unintended bias [3] [4] and other categories of errors. Managers have a difficult time trusting key business decisions to inferences that cannot be explained, especially when bias and other sub-optimal decision making patterns can be embedded within data used for training in non-obvious ways [5] [6]. Even worse, if the output of the application can impact its input, it can maintain a vicious cycle of bias that mires the application in a realm of poor predictive performance indefinitely [7].

With so much at risk, it is time to develop a set of best practices to help guide deployers of AI systems around these potential pitfalls. As a first step in this direction, we propose a framework to understand, quantify, and mitigate sources of bias. First, we survey the potential sources and impacts of bias in AI systems. Then we describe processes that can help companies manage bias to improve their AI outcomes and help provide confidence in the results for themselves, their customers, and regulators.

2. Potential Problems

Bias in AI systems comes from diverse sources, from the algorithm and input attributes chosen to the hidden correlations within training data. We describe three general classes of bias: those related to mapping the business intent into the AI implementation, those that arise due to the distribution of samples used for training (including historical effects), and those that are present in individual input samples.

2.1. Issues with Representing the Goal

The earliest stage for bias to creep into an AI system is when the deployer determines its concrete algorithmic objective from what may be a nebulous goal. For example, the actual goal of a business may be to direct advertising to the potential customers most likely to purchase their product. Since there is no straightforward way to map this into an AI implementation [8], companies must choose the hypothesis, input attributes, and training labels or reinforcement criteria that they deem will best accomplish this goal. For example, a company selling video games might hypothesize that their product would be most likely to sell to young men and look for customers with attributes such as male and aged 15-25.

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.
WWW '19, May 13–17, 2019, San Francisco, USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3317590

2.1.1. Proxy Goals. In the example above, since there is no way to determine the exact likelihood of someone to purchase their product, they may choose a goal of selecting individuals with attributes similar to those of customers that previously purchased a similar product. However, this may not be the right choice when trying to enter new markets or trying to market new product features. Basing the goal choice on historical information without factoring in suitable context will necessarily expose the system to historical bias. For example, people from some regions may simply not have purchased the product because it wasn't previously marketed in their region. Similarly, changes in product features, price, or external trends may render it suitable for customers who had not purchased the previous version.

2.1.2. Feature Selection. The choice of which attributes to include is perhaps the most obvious source of bias from the mapping designers. For example, a university admission system could look at standardized test scores, rank in class, GPA, letters of recommendation, etc. While the end goal may be the same (i.e., predicting success in college), the choice of features used can result in very different decisions. Even seemingly innocuous input features can contain hidden biases [5, 9]. Even more difficult to quantify is bias that arises from features that are not included but could favorably influence the predictions for some people if included. Additional information, such as personal observations from a letter of recommendation, may be harder to map to quantifiable and well-defined features, rendering such useful information unavailable to the algorithm.

2.1.3. Surrogate Data. AI models demand large sets of mathematical input. Training sets must be represented numerically, and because the training sets must be large, these features are limited to those that developers can easily acquire at scale. This may mean that a job screening service uses credit scores rather than letters of recommendation as a surrogate for a feature such as "reliability" [29]. Such mathematical reductions can cause information loss which then biases the problem mapping. Surrogate data can also introduce bias by serving as proxies for restricted input, for example, zip codes can be used as proxy for race, magazine subscriptions for race or gender, and purchasing patterns for medical conditions.

2.2. Issues with Data Sets

In addition to dataset issues that can create problems during the mapping phase, there may be issues with training or production datasets. Creating training sets can be an arduous task. It typically involves preening a large data set [10] and for supervised learning, it includes obtaining labels. For deep learning systems, it may include ensuring rare cases are proportionally overrepresented to give the model adequate opportunity to learn such cases during training. Due to the scale, complexity, and sometimes timeliness of this task, creating a training data set can be the bulk of the effort required in AI systems and is often the source of problems [11]. Training datasets can also be manipulated, rendering the AI algorithm vulnerable [13].

2.2.1. Unseen Cases. Much of the advantage of AI systems are their ability to generalize solutions with robustness to varied input. However, this can become a disadvantage when the system is faced with a class for which it was not trained. For example, a neural network trained to classify texts as German or English will still provide an answer when given a text in French rather than saying "I don't know". Such issues can lead to "hidden" or "silent" mispredictions, which can then propagate to cause additional harm to the business application.

2.2.2. Mismatched Data Sets. If data seen in production differs significantly from that used in training, the model is unlikely to perform well. Extending the point above, commercial facial recognition systems trained on mostly fair-skinned subjects have vastly different accuracies for different populations: 0.8% for lighter-skinned men and 34.7% for darker-skinned women [30]. Even if the model is originally trained on a dataset that matches production use, production data can change over time due to various effects from seasonal changes to external trigger events. Any such change can bring about hidden effects generated by mismatched data sets.

2.2.3. Manipulated Data. Training data can be manipulated to skew the results as was exemplified by the short-lived chatbot Tay, which quickly mimicked the hate speech of its Twitter correspondents [12, 31]. Systems trained on small, public data sets are especially vulnerable to this form of attack. Similarly, data poisoning is a known security challenge for AI systems [13].

2.2.4. Unlearned Cases. Even well-trained models do not have 100% accuracy; indeed, such high accuracy would likely result from overfitting the data and indicate that the model is not likely to generalize well to new cases. As a result, even well-trained models will have classes of samples for which they perform poorly. Studies have shown that facial recognition datasets that do not adequately represent across ethnic groups can cause trained models to display vastly different accuracies across race [14].

2.2.5. Non-Generalizable Features. Due to the practical difficulties in creating large, labeled training sets, model developers may rely on training from well-preened subsets of their expected production data sets. This can result in granting importance to features that are particular to the training set and not generalizable to broader data sets. For example, [15] shows how text classifiers, which were trained to classify articles as "Christian" or "atheist" on standard newsgroup training sets, emphasize non-relevant words like "POST" in making their classifications due to the distributions of those words in the training set.

2.2.6. Irrelevant Correlations. If the training data contains correlations between irrelevant input features and the result, it may produce incorrect predictions as a result. For example, Ribeiro et al. trained a classifier to differentiate between wolves and dogs with images of wolves surrounded by snow and dogs without snow. After training, the model sometimes predicts that a dog surrounded by snow is a wolf [15]. Unlike non-generalizable features, the distribution of irrelevant correlations may not be particular to the training set but may occur in real-world data as well. It may well be that wolves are more likely to be found in snow than dogs. However, it would be incorrect for the feature to impact the prediction; a wolf is still a wolf even when it is in Grandmother's house.

2.2.7. Issues with Using Historical Data. AI systems necessarily learn from past information. Unfortunately, this includes learning human biases contained therein [5] and potentially missing opportunities that emerge from changing environments [9].

2.3. Issues with Individual Samples

We classify issues with individual samples as those that can be seen by examining the data of a single sample. The problem may be with just that sample or endemic to all the samples. This classification is important when the data sets contain personal information such that the entire data set cannot be made publicly viewable, but where individuals may be able to review their own personal data.

2.3.1. Inaccurate Data. Data used for training is often highly curated in order to ensure the model learns effectively. Unfortunately, real-world data is seldom so clean. It can be incomplete or corrupted. For data that is manually entered, it can be entered incorrectly [27]. Data that is automatically collected can have incorrect sources [28].

2.3.2. Stale Data. Data used for both training and production input can be out-of-date. This may particularly be the case when large “dictionaries” are cached for fast access time. For example, credit scores could be downloaded from an external source and stored locally for fast access. Unfortunately, there may be resistance to updating the dataset by developers as it may reset the baseline for ongoing training experiments.

3. Managing Bias

Given the number of potential sources of bias currently known (and more are being discovered as the field matures), it can be daunting to face how to tackle them. Considering the myriad of issues that cause bias to enter the system, we cannot expect a single approach to resolve all of them. Instead, we propose a combination of quantitative assessments, business processes, monitoring, data review, evaluations, and controlled experiments. Before detailing the above stages, we establish some ground rules for processes we want to include.

First, any process to evaluate an AI system for bias must be achievable by those that are not necessarily the primary developer. This is important because the system may need to be understood by non-technical management or evaluated by an auditor or regulator. In fact, since complex models are increasingly available from external sources [16, 17], even the primary developer may not know the model’s inner workings. Therefore, we require that our processes not require any knowledge of the internal workings of the model itself; we treat the entire AI pipeline (including any feature engineering) as a black box and use only the input and output data for evaluation.

Second, transparency of input data is important. This is the only way to verify that the data is accurate and does not contain protected or incorrect data. Even if the data is private, it should be viewable by the person about whom the data is [18]. Methodologies for data versioning, data cataloging, and data tracking and governance are also critical to ensure that the specific dataset used to train a given model can always be identified and examined.

We group the processes into stages at which the process most prominently come into play during the deployment lifecycle of the AI system. However, this is mostly an organizing technique for planning these processes as the stages are likely to overlap and iterate over the lifetime of the application.

3.1. Substantiate Assumptions

During the planning stage of the application, prepare to provide quantitative evidence for the validity of your chosen numerical representations, the hypothesis itself, and the impact of the application on its environment, including its future input.

3.1.1. Substantiate Surrogate Data. When surrogate data is used, it should be accompanied by quantitative evidence that suggests that the surrogate data is appropriate for its intended use in the model. Known limitations of the surrogate data should be documented and presented during reviews of predictions [19].

3.1.2. Substantiate the Hypothesis. Similarly, the model’s intended use and appropriateness for that use should be accompanied by quantitative external evidence that supports it along with any known limitations to the methodology [19].

3.2. Vet Training Data

Training data needs to be vetted for accuracy, relevance, and freshness. Since training data is often curated, it tends to be “cleaner” than inference data. For example, incomplete or ambiguous samples may be eliminated from the set. Unfortunately, the effort required to curate the training data may be at odds with keeping it fresh. On the other hand, production inference data is likely to be fresh but may contain samples that are incomplete or ambiguous.

3.2.1. Avoid Overly Curated Training Data. When considering whether the training set is appropriate for the expected production workload, avoid selecting highly curated sets that perform well on curated evaluation data but are unlikely to perform well in production.

3.2.2. Guard Against Manipulation. Consider whether the training data can be manipulated by outside (or inside) actors. If manipulation cannot be prevented, consider alternative data sources. Suciu et al. present a helpful overview of literature on poisoning machine learning in the context of a generalized model for the capabilities of adversarial agents [20]. Data security practices may also be needed to ensure that inappropriate datasets are not accidentally accessed by those building AI models. In large organizations where teams of data scientists work, such access restrictions can be critical.

3.3. Evaluate for Bias

As part of your model’s evaluation, include the following in addition to standard accuracy metrics.

3.3.1. Utilize Targeted Toolsets. With the growing recognition of issues of bias in AI, there are an increasing number of tools available to help detect and mitigate bias in attribute sets, feature engineering, or the model itself [21].

3.3.2. Validate Predictions. Although we treat AI systems as a black box, there are tools available that can determine which input features determined the resulting prediction [15, 22]. Such predictions can be validated by providing a qualified judge (who does not a priori know the AI system's prediction) with the same feature values and comparing the judge's determination with the AI's prediction. This both validates and provides explanations for predictions. Such an evaluation process can help identify both non-generalizable features and irrelevant correlations.

3.3.3. Match Training to Production Needs. It is important to monitor the distribution of input data to see whether production data is consistent with the data used in training. Input data should be monitored for anomalous cases that differ substantially from those seen in training. Input streams should also be monitored to ensure that the distribution of data seen in production does not stray from the distribution anticipated by the training data set.

3.4. Monitor Production Data

AI accuracy is based on its evaluation data set. When the input seen in production does not match this, the system's performance should be flagged and treated with skepticism.

3.4.1. Detect Incomplete Data. Production AI systems should actively monitor for incomplete data. This is particularly true when the system is trained on cleaned data sets but then receives a wider range of samples in production. For example, an employer recruiting program may use credit scores to train a filter to screen candidates and confuse someone with no credit history as someone with a low credit history.

3.4.2. Detect Data Divergence. Production input monitoring should evaluate whether the input data received for inferences matches the expected distribution anticipated by the training data. If not, the model should be re-trained with fresher data to match the production workload [23].

Note that this is similar to the need to match training data to production needs that we discussed earlier. The first occurs during the planning of the training set and the second occurs in monitoring production data to detect when assumptions made during the training phase no longer hold.

3.4.3. Detect Uncommon Cases. Inputs with unique feature distributions or low confidence should be detected and ideally flagged for external labeling and inclusion into future training sets. Chen et al. have shown that supplementing training sets with additional data can be an effective strategy for reducing bias [24].

3.4.4. Refresh Input Data Regularly. Input data sources should be refreshed in a timely manner. This includes data used for training as well as any stored data sets that are included as features for production inference.

3.5. Create Supportive Processes

In addition to the data science and engineering aspects of managing bias, there also needs to be business practices in place to enable reviews and emphasize testing new directions.

3.5.1. Enable Input Data Review and Correction. Individuals should be able to view data that is about them personally. There should be a process in place for them to challenge and correct this data [18]. Ideally, there should also be processes for individuals to provide alternate information that directly speaks to the goal (and not just the proxy goals) of the system and have that additional information considered by a human reviewer.

3.5.2. Create Mechanisms to Overcome Historical Bias. Adding some randomization to outcomes can be used to avoid consistent discrimination from historical data [25]. Similarly, A/B experiments can be used that drop some features. Ideally, the results of such predictions can be fed back into the system to produce continuous improvement.

3.6. Quantify Feedback Loops

In cases for which the AI predictions can impact its own future input, it is important to ensure that this impact is quantified by comparing against a suitable control or comparing the results against external evidence. For example, algorithms that predict crime based on police reports may cause more police to be deployed to the site of the predictions, which in turn may result in more police reports [7]. Note that not all feedback loops are considered negative; algorithms that return search results may return better search results when less common search terms are used, which may result in users learning to enter less common search terms. However, even in positive cases, it is important for the designers to understand the impact of feedback loops.

4. Related Work

While there exists robust literature of potential problems with bias in AI systems, there is considerably less available for how to manage this bias, especially across the wide spectrum of potential causes of bias. A number of principles documents and primers for algorithmic accountability have been published by researchers [32] and groups such as ACM [33, 34], World Wide Web Foundation [35] AI Now [36], and Data and Society [37]. Other researchers have proposed taxonomies of bias that serve as an excellent starting point towards managing that bias [38, 39, 40] Cramer et. al describe their experience with translating these principles documents and taxonomies into concrete processes within their organization, Spotify [26]. We similarly share the goal of mapping research literature into practical processes for industry teams and focus on processes that can be integrated into existing product cycles including substantiating surrogate data, vetting training data, and validating predictions. Cramer et al. point out that in early product development data sets may be limited and thus teams may be making decisions in the presence of skewed information. Some of the techniques we propose for detecting data divergence can help with this by actively monitoring and reacting when production data varies substantially from the assumptions made during training.

5. Summary

Many data sets contain bias that inevitably impacts AI systems that use that data. Even careful cleansing of protected and proxy data will not completely remove the results of historical bias contained within the data. However, putting in place a combination of processes can mitigate the impact of bias on current predictions and reduce its impact over time. We advocate a combination of processes:

Table 1: Issues that cause bias and recommended mitigation actions.

Issue type	Issue	Mitigations
Representing the problem	Proxy Goals	• Substantiate the hypothesis with external quantitative evidence. • Evaluate impact of predictions against external metrics.
	Feature Selection	• Utilize targeted toolsets to help identify hidden bias. • Validate prediction reasons as part of model evaluation. • Evaluate unlearned cases for clues to missing attributes.
	Surrogate Data	• Substantiate surrogate data noting known limitations. • Enable an input data review process to help catch non-obvious limitations.
Datasets	Unseen Cases	• Detect and review unexpected input feature patterns.
	Mismatched Data Sets	• Avoid overly curated training data. • Detect when input data diverges from distributions anticipated during training.
	Manipulated Data	• Consider how training data and data used as input could be manipulated and enable safeguards or use alternative sources.
	Unlearned Cases	• Review the cases that were not learned during training as part of model evaluation.
	Non-generalizable Features	• During model evaluation, determine which features were used to make predictions and compare them against those used by an external judge.
	Irrelevant Correlations	
	Issues with Historical Data	• Utilized targeted toolsets that help remove historical bias. • Supplement training sets with under-represented samples. • Utilize randomization and A/B experiments to explore new outcomes.
Individual Samples	Inaccurate Data	• Detect incomplete data. • Enable an input data review process to help catch and correct incorrect input.
	Stale Data	• Refresh cached data sets regularly.

- Data monitoring to detect anomalies and determine whether the training set is appropriate.
- Quantitative analysis to justify surrogate data and account for feedback loops.

- A review process that ensures input correctness.
- An evaluation process that verifies the predictions are grounded in reasonable feature values.
- Mechanisms to explore new outcomes and incorporate these outcomes into future predictions to overcome historical bias.

Table 1 maps the problems described in Section 2 to the mitigation actions described in Section 3.

Reviews of AI systems for bias should ensure all of the above are implemented. As AI systems become more competitive, mitigating bias may become more necessary not only to avoid liability but to explore opportunities that may otherwise be missed. As awareness of the potential harmful impacts of bias in AI algorithms grows, we can encourage companies to mitigate the bias in AI systems by emphasizing the benefits of doing so such as increasing competitiveness and avoiding hazards and by outlining practical management processes that can be followed to actively identify and reduce that bias.

REFERENCES

[1] Witten, H., Frank, E., and Hall, M.A. (2011). Data Mining: Practical Machine Learning Tools and Techniques, 3rd ed., Morgan Kaufmann Publishers Inc., San Francisco, CA.

[2] McKinsey Global Institute (2017). “Artificial Intelligence: The Next Digital Frontier?”

[3] Goodman, B. and Flaxman, S. (2016). “European Union regulations on algorithmic decision-making and a 'right to explanation'”. ICML Workshop on Human Interpretability in Machine Learning, New York, NY.

[4] Kahn, J. (2018). “Artificial Intelligence Has Some Explaining to Do”. *Bloomberg Businessweek,* 12 Dec 2018.

[5] Bolukbasi. T., Chang, K., Zou, J., Saligrama, A., and Kalai, A. (2016). “Man is to computer programmer as woman is to homemaker? Debiasing word embeddings”. *Proceedings of the 30th International Conference on Neural Information Processing Systems.*Barcelona, Spain.

[6] Angwin, J., Larson, J, Mattu, S., and Kirchner, L. (2016). “Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks”. ProPublica, 23 May 2016.

[7] Ensign, D., Friedler, S.A., Neville, S., Scheidegger, C. and Venkatasubramanian, S. (2018). “Runaway Feedback Loops in Predictive Policing”. *Proceedings of Machine Learning Research.*

[8] Hao, K. (2019). “This is How AI Bias Really Happens -— and Why It’s so Hard to Fix”. MIT Technology Review, 4 Feb 2019.

[9] Dastin, J. (2018). “Amazon scraps secret AI recruiting tool that showed bias against women”. Reuters Business News, 10 Oct 2018.

[10] Del Balso, M. and Hermann, J. (2017). “Meet Michelangelo: Uber's Machine Learning Platform”. Uber Engineering, 5 Sep 2017.

[11] Baylor. D. et al. (2017). “TFX: A TensorFlow-Based Production-Scale Machine Learning Platform”. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Halifax, Nova Scotia, Canada.

[12] Vincent, J. (2016). “Twitter taught Microsoft’s AI chatbot to be a racist asshole in less than a day”. The Verge, 24 Mar 2016.

[13] Bursztein, E. (2018). “Attacks against machine learning - an overview”. https://elie.net/blog/ai/attacks-against-machine-learning-an-overview/.

[14] Buolamwini, J. (2019). “Aritificial Intelligence Has a Problem with Gender and Racial Bias”. TIME, 7 Feb 2019.

[15] Ribeiro, M.T., Singh, S., and Guestrin, C. (2016). “’Why Should I Trust You?’: Explaining the Predictions of Any Classifier”. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Francisco, CA.

[16] Google. "TensorFlow Hub," https://www.tensorflow.org/hub.

[17] H2O. “Open Source Leader in AI and ML”. h2o, https://www.h2o.ai/.

[18] European Commission. “Rights for Citizens - European Commission”. https://ec.europa.eu/info/law/law-topic/data-protection/reform/rights-citizens_en.

[19] Board of Governors of the Federal Reserve System (2011). “The Fed - Supervisory Letter SR 11-7 on Guidance on Model Risk Management”. https://www.federalreserve.gov/supervisionreg/srletters/sr1107.htm.

[20] Suciu, O., Marginean, R., Kaya, Y., Daume. H. III, and Dumitras, T. (2018). “When Does Machine Learning FAIL? Generalized Transferability for Evasion and Poisoning Attacks”.*27th USENIX Security Symposium*, Baltimore, MD.

[21] IBM. “AI Fairness 360”. http://aif360.mybluemix.net/.

[22] H2O. “Interpreting a Model — Using Driverless AI”. https://www.h2o.ai/wp-content/uploads/2017/09/driverlessai/interpreting.html.

[23] Ghanta, S. and Talagala N. (2018). “MLOps Health: Taking the Pulse of ML in Production”. ITOpsTimes, 10 July 2018.

[24] Chen, I.Y., Johansson, F.D., and Sontag, D. (2018). “Why Is My Classifier Discriminatory?”.*32nd Conference on Neural Information Processing Systems,* Montreal, Canada.

[25] Kroll, J., Huey, J., Barocas, S, Felten, E., Reidenberg, J., Robinson, D., and Yu, H. (2017). “Accountable Algorithms”. *University of Pennsylvania Law Review,* vol. 165, pp. 633-705.

[26] Cramer, H., Garcia-Gathright, J., Springer, A, and Reddy, S. (2018). “Assessing and addressing algorithmic bias in practice”. *ACM Interactions,* vol. XXV, no. 6, pp. 58-63.

[27] Bier, D. (2017). “E-Verify Has Delayed or Cost Half a Million Jobs for Legal Workers”. CATO Institute, 16 May 2017. https://www.cato.org/blog/e-verify-has-held-or-cost-jobs-half-million-legal-workers.

[28] Suhartono, H., Levin, A., and Johnsson, J. (2018). “Why Did Lion Air Flight 610 Crash? New Report Details Struggle”. Bloomberg, 27 Nov 2018. https://www.bloomberg.com/news/articles/2018-11-27/lion-air-pilots-struggle-detailed-in-preliminary-crash-report.

[29] 0'Neil, C. (2016). Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. Crown. New York, NY.

[30] 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*, New York, NY.

[31] Buranyi, S. (2017). “Rise of the racist robots - how AI is learning all our worse impulses”. The Guardian, 8 Aug 2017. https://www.theguardian.com/inequality/2017/aug/08/rise-of-the-racist-robots-how-ai-is-learning-all-our-worst-impulses.

[32] Nicholas Diakopoulos and Sorelle Friedler (2016). How to Hold Algorithms Accountable, MIT Technology Review, November 17, 2016. https://www.technologyreview.com/s/602933/how-to-hold-algorithms-accountable/

[33] Garfinkel, S., Matthews, J., Shapiro, S., and Smith, J. (2017). Toward Algorithmic Transparency and Accountability. Communications of the ACM. Vol. 60, No. 9, Page 5, Sept. 2017, 10.1145/3125780.

[34] ACM US Public Policy Council (USACM) (2017). Statement on Algorithmic Transparency and Accountability. January 12, 2017. https://www.acm.org/binaries/content/assets/public-policy/2017_usacm_statement_algorithms.pdf

[35] World Wide Web Foundation (2017). Algorithmic accountability report, 2017. https://webfoundation.org/docs/2017/07/Algorithms_Report_WF.pdf

[36] Reisman, D., Schultz, J., Crawford, K., and Whittaker, M. (2018). Algorithmic impact assessments: A practical framework for public agency accountability. AI Now institute. Apr. 2018. https://ainowinstitute.org/aiareport2018.pdf

[37] Caplan, R., Donovan, J., Hanson, L., and Matthews, J. (2018). Algorithmic accountability primer. Data & Society. Apr. 2018. https://datasociety.net/wp-content/uploads/2018/04/Data_Society_Algorithmic_Accountability_Primer_FINAL-4.pdf

[38] Friedman, B. and Nissenbaum, H. (1996). Bias in computer systems. ACM Trans. Inf. Syst. 14, 3, 330–347.

[39] Olteanu, A., Castillo, C, Diaz, F., and Kiciman, E. (2016). Social data: Biases, methodological pitfalls, and ethical boundaries.http://dx.doi.org/10.2139/ssrn.2886526

[40] Baeza-Yates, R. (2016). Data and algorithmic bias in the web. Proceedings of the 8th ACM Conference on Web Science. ACM. New York. https://doi.org/10.1145/2908131.2908135