Weapons of Math Destruction Revisited: Addressing Opacity, Regulation, and Fairness in Machine Learning

Francesca Albio, Petra Kocsis, Raluca Dinu, Zita Rátkai

Amsterdam University College

Machine Learning

Breanndán Ó Nualláni

10 December 2024

1. Introduction

Cathy O’Neil is an American mathematician, data scientist, and author. In her New York Times bestseller, *Weapons of Math Destruction (2016),* she discusses the issues surrounding machine learning systems, analyzing the hidden dangers, especially in the fields of criminal justice, finance, and hiring. She raises different problems including opacity, lack of regulations, reinforcement of discrimination and also discusses some of the solutions that could contribute to making these systems more fair and ethical. This report aims to present the key themes from O’Neil’s analysis, starting with the opaqueness of such systems and their consequences. It then explores the issues raised by accountability and the absence of regulations and how machine learning systems reinforce discrimination by amplifying biases present in their data or design. Ultimately, it discusses the proposed solutions and the current research aimed at achieving fair outcomes. By examining these aspects, this report provides a critical perspective when it comes to challenges posed by machine learning. Through O’Neil’s arguments and additional research, it emphasizes the need for transparency, accountability and fairness to ensure ethical use of such systems.

2. Opaqueness of Machine Learning Models

One core problem of machine learning described by Cathy O'Neil in Weapons of Math Destruction as: the opacity of models, also known as "black boxes." These systems are defined by their ability to make predictions and complete classification tasks without providing any clear explanation of how or why each decision was reached. This lack of transparency is not just a technical limitation, but also raises profound ethical and societal concerns, particularly when implemented in areas such as criminal justice, hiring, and finance, where decisions carry significant consequences. O’Neil argues that when individuals that are affected by these decisions cannot understand or challenge them, it creates a dangerous imbalance of power and accountability (O’Neil, 2016).

One of the main reasons for this opacity has to do with the computational complexity of contemporary machine learning models. Fernández et al. (2016) point out that interpretability is one of the major challenges in computational intelligence systems, especially when they are developed to handle large volumes of data. Their results show how models developed for big data often become so complex, that even their developers have difficulty understanding their internal mechanism. Similarly, López et al. (2015) highlights the challenges of making machine learning systems interpretable. They note that as machine learning systems handle more variables and imbalanced datasets, their complexity progressively increases, making their decision making ever harder to understand.

Their complexity is often the result of a tradeoff between performance and interpretability. According to Triguero et al. (2015), machine learning systems aim to enhance efficiency and accuracy when handling large volumes of data, which however comes at the expense of model transparency. They explained that dataset imbalances may affect a model's predictions and introduce biases in the system, which are quite difficult to trace and correct because of the system's lack of interpretability. This tradeoff reflects a larger challenge that O’Neil identified as: the pursuit of higher accuracy often undermines the goals of fairness and accountability (O’Neil, 2016).

The consequences of this opacity go far beyond technical debates. O'Neil claims that in decision-making contexts, opaque machine learning models used in criminal sentencing, hiring, education, and other areas gradually undermine individuals and escape accountability for it. When these systems are used to provide validation for decisions without surveillance, they create a self-confirming cycle of authority. For example, when a biased model decides who can or cannot be hired, those left out have little chance of understanding, let alone opposing to, the rationale behind the decision. This lack of contestability further reinforces systemic inequities and discrimination, one of the main tools of "Weapons of Math Destruction" according to O'Neil.

Ultimately, the opaqueness of machine learning models sets up a dual threat: the technical limitations that prevent its interpretability and the ethical risks that increase inequality. As Fernández et al. (2016) and López et al. (2015) point out, this problem will be further provoked by the increasing complexity of systems developed to process large-scale and imbalanced data. The prioritization of performance over interpretability, alongside the lack of regulatory supervision, underlines O'Neil's (2016) critique of machine learning as an opaque, unregulated force with potentially dangerous consequences for society. Therefore, there is a necessity for a shift in the balance between efficiency and transparency, so that the machine learning systems, while remaining accurate, would be held accountable and fair.

3. Lack of Regulation and Accountability (≈ 450 words)

Cathy O’Neil addresses the issue of lack of regularization and accountability in machine learning systems. These systems operate in absence of a meaningful oversight which creates a dangerous environment where biased and opaque algorithms can worsen societal problems. To ensure ethical and equitable functioning of these, regulation and accountability are essential, however, achieving this is still a challenge.

One major issue in machine learning are models evolving faster than the policies to govern them. Traditional laws such as anti-discrimination law are not equipped to address the unique challenges these innovations bring. These problems lead to regulatory gaps which make it harder to ensure fairness as well as ethical decision making.

The evolution of these models also creates accountability gaps which are further complicated by the structure of machine learning development. Developers, data providers and users all contribute to the model outcome, but this poses an issue when it comes to assigning responsibility. The issues become more complex in globalized systems where machine learning models are created in one country and deployed in others, making it unclear whose laws they should follow.

An obstacle in accountability is the complexity and opacity of machine learning systems. Many companies claim their algorithms as intellectual property to avoid disclosing how the model functions, even when these algorithms make life altering decisions. This makes it impossible for audits to reveal unfair practices or biases. Moreover, the complexity of these algorithms often makes them unintelligible to the general public, creating a gap between the developers who design the systems and the individuals or groups who are in charge of questioning or challenging their outcomes.

The lack of regularization and accountability poses an issue when it comes to societal trust in machine learning systems as well. Mehrabi et al. (2021) explain how the lack of transparency intensifies public skepticism when it comes to technology, especially in areas like healthcare and criminal justice where critical decisions are made. For example, a healthcare algorithm based on biased historical data might exacerbate inequalities in treatment access and marginalize vulnerable populations.

In her book, O'neil advocates for algorithmic transparency and accountability. She proposes methods to address these issues, focusing first in mandatory audits of machine learning systems. These audits would allow experts such as researchers, journalists, civil society organizations to evaluate how these models work and whether they harm certain groups. Moreover, O'neil proposes full disclosure of the data and the logic behind the algorithms, which provides crucial transparency. By proceeding this way the companies can be held accountable for the outcomes whether they are harmful or not.

Furthermore, O'neil suggests a regulatory framework that can keep up with the development of machine learning. Fernandez et al. (2014) emphasize that by arguing that big data systems are more integrated in decision making and a lack of regulatory oversight can lead to harmful outcomes such as discrimination or exclusion.

4. Reinforcement of Discrimination (≈ 450 words)

Cathy O'Neil's analysis discusses the pervasive issue of machine learning models reinforcing discrimination. Biases embedded in the data or algorithms can lead to systematic inequalities, creating a feedback loop that amplifies existing disparities. In machine learning systems that influence important decisions, such as criminal justice, hiring, and finance this phenomenon is extremely concerning.

Data often reflects societal biases and machine learning models trained on biased data are very sensitive. Models are likely to perpetuate harmful biases if the data they trained on is imbalancedcontaining skewed class distributions or underrepresenting certain groups. Triguero et al. (2015) highlight this issue in their work on imbalanced data classification, where traditional machine learning algorithms struggle to predict minority class instances correctly. The model favors the majority class and marginalizes already disadvantaged groups, leading to unfair outcomes.

Triguero et al. provide us with possible solutions to this issue, evolutionary undersampling methods to address class imbalances in big data. These techniques reduce the overrepresentation of majority class instances, resulting in a more balanced dataset. This method improves classification accuracy for minority classes, however, it underscores the fragility of machine learning models when trained on skewed data. The fact that data distributions must be changed to achieve fairness implies that algorithms are not neutral by nature but are instead influenced by the inputs and design decisions made.

Similar to this, López et al. (2015) investigate how fuzzy rule-based classification systems can handle unbalanced datasets. The systems they observed used linguistic rules to create more understandable and equitable decision-making frameworks. They found that even these advanced methods cannot fully remove the risk of discrimination if the data reflects underlying societal inequalities. Without deliberate efforts to balance for bias during data preprocessing and algorithm design, machine learning models are prone to perpetuate systematic discrimination.

O'Neil uses a real-world example to further her point in the use of predictive policing algorithms. The systems she discusses are trained on historical crime data, which usually disproportionately targets marginalized communities due to already existing societal biases in law enforcement. These algorithms direct police to neighborhoods of these marginalized groups, increasing the arrest rates for these areas, therefore reinforcing the initial biases. This is a self-perpetuating feedback loop and the discrimination becomes more and more embedded in the system over time.

Triguero et al. and López et al. made points that align with O'Neil's concerts, proving that biases in machine learning aren't just technical errors but reflections of broader societal inequities. Without intentional interventions like fairness constraints, diversifying training data, and designing algorithms with equity in mind, these models will only further replicate and worsen discrimination. There's an urgent need for ethical guidelines and regulations to ensure that these models serve all members of society equally.

5. Solutions proposed by O'Neil and current research

In addressing the problems posed by Weapons of Math Destruction, O'Neil (2016) advocates a multi-pronged set of reforms that begin with transparency, accountability, and a values-driven orientation toward how predictive models are built and used. First, data scientists themselves must acknowledge their societal role by adhering to a professional code of ethics that prioritizes fairness and honesty over blind profit-seeking. Beyond personal responsibility, she stresses the need for systematic auditing of influential algorithms. Independent researchers, journalists, and public-interest organizations should regularly examine how these models operate, assess their hidden biases, and publish their findings openly. Such audits would help the public understand when models cause harm and hold institutions accountable. On the regulatory front, O'Neil suggests extending and strengthening laws to cover emergent scoring practices—such as online credit evaluations and personality screenings—and to ensure that protected categories are not discriminated against through indirect data proxies. Consumer protections should guarantee that individuals know when and how their data is used, and provide avenues to challenge inaccuracies. Additionally, health and employment protections must be updated to prevent the exclusion of people based on "predicted" health risks or conditions revealed by data analytics. A critical element of reform involves rethinking how success is defined in algorithmic systems. Rather than focusing solely on efficiency or profit, metrics must include fairness and social welfare. Additionally, O'Neil encourages the use of data modeling for socially constructive purposes. Models can identify where resources are needed most, highlight discriminatory patterns for correction, and guide policy decisions that genuinely serve the public interest. When models are designed to support human judgment rather than replace it, and when they incorporate meaningful feedback loops, they can evolve to become less harmful and more just.

Building on O'Neil's emphasis on transparency and accountability, subsequent theoretical work has delved deeper into the complexity of implementing fair algorithmic decision-making processes. For instance, Kleinberg, Mullainathan, and Raghavan (Kleinberg et al., 2016) investigated scenarios in which multiple reasonable fairness constraints cannot all be met at once. Their analysis shows that distinct notions of fairness, such as ensuring predictive accuracy across groups or balancing error rates, often prove incompatible. A key reason for this incompatibility lies in the underlying distributions of the groups being compared. If two groups differ in their base rates—the proportion of individuals actually exhibiting the outcome of interest—then meeting all fairness criteria simultaneously is mathematically impossible unless the model achieves near-perfect prediction or the groups share identical base rates. Conditions like calibration, which require predicted probabilities to match observed outcomes, and parity in error rates, which demand equal misclassification patterns across groups, inherently pull the model's predictions in conflicting directions. Thus, this insight does not negate the importance of striving for just and equitable systems, but it does mean that stakeholders must confront difficult trade-offs, carefully selecting which fairness criteria align best with their ethical and policy objectives, given the fundamental, unavoidable tensions when working with heterogeneous populations.

Beyond the complexities identified by Kleinberg and colleagues, more recent surveys in machine learning have further clarified the landscape of bias and fairness techniques. Mehrabi et al. Mehrabi et al. (2021) present an extensive review of the bias sources and fairness strategies employed across various AI domains, including natural language processing and deep learning. They discuss how methods such as causal inference, representation learning, and adversarial training can mitigate unfairness, while also highlighting persistent challenges, such as reconciling different notions of fairness and detecting hidden biases. Their comprehensive overview shows that, despite significant progress, designing fair AI systems remains a multifaceted endeavor requiring careful consideration of context, domain-specific constraints, and ethical trade-offs.

Recent work by Selbst et al. (2019) offers another perspective on the difficulty of achieving fair outcomes in complex, real-world environments. They argue that fairness efforts in machine learning often rely on foundational computer science principles—like abstraction and modular design—to define and implement fairness criteria. However, these very principles risk distorting or oversimplifying the social contexts in which models operate. According to their analysis, attempting to solve fairness issues strictly within the confines of technical abstractions can conceal critical aspects of social reality, including nuanced power dynamics, shifting cultural norms, and the institutional practices that shape how data are produced and used. The authors identify several pitfalls—such as failing to consider the broader decision-making ecosystem or overlooking the ways that tools can be repurposed—that can render technical fairness interventions ineffective or even harmful. They encourage designing processes, rather than fixed solutions, that explicitly integrate social actors and acknowledge that fairness is not a static property of technology alone, but a property of entire sociotechnical systems.

6. Conclusion

In conclusion, Cathy O'Neil's (2016) book ‘Weapons of Math Destruction’ discusses some of the challenges with machine learning models, including their opacity, lack of regulation, reinforcement of discrimination, and need for improvements of the systems. The book uncovers how algorithms have come to possess unchecked power in criminal justice, hiring, and finance, and raises its ethical and societal red flags. These "black box" systems lack transparency, creating imbalances of power and perpetuating inequality.

O'Neil argues that this performance-interpretability tradeoff is highly problematic, since it makes the decisions of an algorithm untraceable. Fernández et al. (2016) and López et al. (2015) indicate how scaling big data systems decreases transparency, while Triguero et al. (2015) raised concerns about amplifying biases.

In the absence of regulatory frameworks, non-transparent and prejudiced models work unchecked, impacting public trust and causing harm in important areas like health care and justice. O’Neil further criticizes how systemic discrimination is reinforced by these models: imbalanced datasets create feedback loops that make the marginalized disadvantaged.

Triguero et al. (2015) and López et al. (2015) highlight that inequities will remain unless constraints to fairness are created. Due to these challenges, O'Neil calls for the implementation of transparency, accountability, and fairness: audits, ethical codes, and regulation in machine learning systems. Other scholars, such as Kleinberg et al. (2016) and Selbst et al. (2019), stress the complexity of achieving fairness, pointing to its trade-offs and socio-technical nuances.

It is important for the improvement of the systems to continue refining fairness techniques and encourage interdisciplinary collaboration. This involves the engagement of stakeholders, the adaptation to evolving norms, and the understanding of social dynamics as part of machine learning systems over time, to remain ethically responsible and equitable.

References

Fernández, Alberto, et al. *A View on Fuzzy Systems for Big Data: Progress and Opportunities*. Vol. 9, no. Supplement 1, 26 Apr. 2016, pp. 69–69, https://doi.org/10.1080/18756891.2016.1180820. Accessed 8 June 2023.

---. "Big Data with Cloud Computing: An Insight on the Computing Environment, MapReduce, and Programming Frameworks." *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 4, no. 5, Sept. 2014, pp. 380–409, 150.214.190.154/sites/default/files/ficherosPublicaciones/1810_2014-WIRES-Fernandez_etAl-Big_Data_w_Cloud_Computing.pdf, https://doi.org/10.1002/widm.1134.

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent Trade-Offs in the Fair Determination of Risk Scores. ArXiv:1609.05807 [Cs, Stat]. https://arxiv.org/abs/1609.05807

López, Victoria, et al. "Cost-Sensitive Linguistic Fuzzy Rule Based Classification Systems under the MapReduce Framework for Imbalanced Big Data." *Fuzzy Sets and Systems*, vol. 258, Jan. 2015, pp. 5–38, https://doi.org/10.1016/j.fss.2014.01.015. Accessed 29 Aug. 2020.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A Survey on Bias and Fairness in Machine Learning. ACM Computing Surveys, 54(6), 1–35. https://doi.org/10.1145/3457607

O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown.

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and Abstraction in Sociotechnical Systems. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 59–68. https://doi.org/10.1145/3287560.3287598

Triguero, Isaac, et al. "Evolutionary Undersampling for Imbalanced Big Data Classification." *CiteSeer X (the Pennsylvania State University)*, 1 May 2015, https://doi.org/10.1109/cec.2015.7256961. Accessed 14 Aug. 2023.