**Beyond Transparency: Algorithmic Bias, Opacity, and the Right to Understand**

## Introduction

In *Algorithmic Bias: On the Implicit Biases of Social Technology*, Gabbrielle M. Johnson points out important commonalities between algorithmic biases and human biases. They share a similarity in formation—both built upon the flaw of human reasoning, a point I will discuss in after sections. Besides, algorithmic bias is often a result reinforced, or even imposed by human bias; it reflects human bias, and as such, requires social—not merely technological—solutions.

For instance, the "Automating Report 2020" shows that sixteen of the sixteen investigated European countries deployed biased AI systems in public policy or surveillance with inadequate oversight (Fischer et al.). To name a few other scenarios, algorithms weaponize exams against low-income students through postcode-based grading in the UK, systematically target the poor via welfare surveillance in the Netherlands, and normalize flawed facial recognition for mass data harvesting—all under the veneer of "neutral" governance in Belgium, warning us of the prevalence and severity of algorithmic bias.

The essay will first examine algorithmic bias and its defects, as well as algorithmic opacity and how it shields algorithmic bias. Then, the essay will scrutinize how the lack of legal regulations reinforces the issues and evaluate current resolutions. Ultimately, the essay will discuss several key points to resolve algorithmic bias.

Algorithmic bias, shielded by algorithmic opacity and the lack of legal regulation, not only stems from a common origin with human bias in its formation, but also acts as a direct projection of human bias. Hence, in addressing the issue, I argue that there are no purely algorithmic solutions - several attempts to address algorithmic bias have different deficiencies, and thus the key lies in increasing algorithmic intelligibility through the hand of legal regulations.

## Algorithmic Bias - the Proxy Problem and the Flaw of Human Reasoning

According to Gabrielle M. Johnson, algorithmic biases are biases that develop, unintentionally by their programmers, from social patterns reflected in their training data (1). In the initial stage of algorithmic bias, the calculation process includes prejudiced factors, e.g., race, engendering biased results, e.g., the probability of an African American committing a

crime is higher. Today, most programmers have intentionally removed biased factors, yet algorithmic bias largely remains.

According to Johnson, the non-eliminability of algorithmic bias is due to so-called "proxy attributes": "seemingly innocuous attributes that correlate with socially sensitive attributes, serving as proxies for the socially-sensitive attributes themselves" (1). For instance, machines discover that the elderly often have a strong degree of trust in Fox News compared to the younger generation, along with a negative correlation between age and computer skills. For machines, each individual is the combination of innumerable tags and they analyze the correlation between tags based on past data and experiences. Naturally, the fact that there is a strong correlation between trust in Fox News and age would encourage, or even push algorithms to give the answer that whoever trusts Fox News more has an older age. Similarly, the result that whoever is older is worse at computers is achieved. Combining the results, computers eventually show biases - Fox News fans are bad at computers - without including explicit biased factors. This is the so-called Proxy Problem.

To look into the reasoning process, datafication turns each individual into a series of discrete tags that are yet intricately connected; they are then analyzed for patterns—emphasizing calculability over complexity and comprehensiveness (Boyd and Crawford). The fundamental issue lies in the fact that machines fail to pick out the effect of biased attributes on seemingly innocuous attributes. Therefore, as long as the reasoning logic built on "finding correlation based on past data" remains, the simple removal of sensitive factors cannot eliminate algorithmic bias.

The "problematic" algorithmic logic can be traced back to how we humans think. The philosopher Immanuel Kant wrote in *the Critique of Pure Reason* that "[h]uman reason is by nature architectonic, i.e., it considers all cognitions as belonging to a possible system" (502). In an attempt to judge new scenarios out of the "architecture", people seek correlations between them and known understandings, a process of which is named induction. However, the actual relationship between them is far more complicated: we tend to boldly consider causal, probabilistic relationships as deterministic ones. This misinterpretation of random relationships as deterministic ones is the root of bias. Hence, "bias exists everywhere induction does" (Johnson 3). In this sense, bias is built into reason itself. The exact reasoning process we embed into machines brings the exact formation of biases into them. Algorithmic biases persist due to the scope of implicit human bias in our social world, and such results always appear even after eliminating sensitive factors.

Unfortunately, most people don't realize the relationship and aim to minimize human bias with algorithms (Fry). For instance, in 2007, programmers in Washington (who graduated from Princeton) designed a system that graded teachers based on their students'

scores in annual examinations to weed out the low-performing teachers and enhance education experience. It looked great: a way to keep those with great teaching skills and eliminate those who did nothing but ingratiate themselves with administrators, which would reduce human bias. However, the system has three fatal issues. Firstly, it uses over-simplified factors with a small database to measure students' performance of learning - a very complicated process that would need far more other factors beyond scores to judge. In Chinese, the phrase "管中窥豹" - to glimpse a leopard through a bamboo tube - emphasizes how partial and limited the perspective of the algorithm is. Secondly, it is self-perpetuating: "Statisticians use errors to train their models and make them smarter" (Fry 5). As time passes, the bias would only become more and more ineradicable. Thirdly, the process of the judgment is obscure, preventing people from pointing out issues since they cannot even be explained. Simply repeating "the system said" acts as a shield to the attempts to address the issue (Eubanks 10).

**Algorithmic Opacity - the Shield of Bias and the Loophole in Law**

The third point above about obscure judgment is correlated with algorithmic opacity. In The Bl*ack Box Society*, Professor Pasquale phrases opacity as a black box. On the one hand, a black box refers to an imperceptible recording device supervising our lives without consent, representing sharply how algorithms are violating our privacy; on the other hand, it stands for an unintelligible process in which only inputs and outputs are shown (Pasquale 12). We have pointed out that the Proxy Problem is one major issue in the judgment process of machines, while the actual process remains largely in shadow. Opacity acts as a one-way mirror between users and algorithms, from which the latter can see the former while we users, like suspects interrogated by police, have little information and knowledge about the complex algorithmic system.

If we address algorithmic bias as the source of the problem, then algorithmic opacity is the shield protecting it from the public, by the hand of law. It prevents those who have their rights and privacy infringed upon from calling to account. Considering huge disparities in scale between individuals and enterprises, victims already have a natural fear of the complex algorithms and often powerful companies behind them. What's worse, even with enough courage, they have no way to start due to the unintelligibility of algorithms - "We cannot understand, or even investigate, a subject about which is unknown" (Pasquale 1). It is like walking into an enormous room filled with eternal blackness, trying to find and kill the dragon hidden in a corner. Even the smartest and most courageous hero would not be able to accomplish the mission.

Due to the omission or even intentional misguiding of law, algorithmic opacity is somehow reinforced and protected by legal systems. In the protection of enterprises, the Trade Secrets Act (18 U.S.C. 1905) provides criminal penalties for the theft of trade secrets and other business identifiable information (Office of Privacy and Open Government), making it nearly impossible to access the database of algorithms; while "[t]here is no comprehensive national privacy law in the United States" (DLA Piper). Everything we do online is recorded with little to no protection from privacy law, yet the world of commerce receives incommensurate legal incline. Laws are meant to protect the disadvantaged, yet now they reinforce such inequality (France).

Fortunately, the loophole in law is noticed and certain parts of the world are taking action. In April 2021, the European Commission proposed the first EU AI law, the main content of which classifies artificial intelligence into three levels based on their risks. AIs considered with unacceptable risks are directly banned; EU will assess high-risk AI systems before they the market and through their life-cycle; while other AIs not classified as high risks, like ChatGPT, will have to meet transparency requirements (European Parliament). This is humankind's first step toward regulating algorithms using the hand of laws. It is a small yet meaningful act. However, the law still has room for improvement. AI systems considered with no major risks are only regulated by transparency requirements, which indeed is not sufficient in addressing the issue. For instance, a firm can send users a document with 30 million pages in response to information, leaving potential investigators looking for a needle in a haystack (Pasquale 6). In this case, while seemingly reassuring transparency, complexity is instead weaponized by firms to stay in the dark. Therefore, simply increasing transparency is not sufficient. Laws would have to state the requirements for the intelligibility of algorithms clearly, not just endowing each individual the right to investigate, but also keeping the path of investigation flat and clean.

**Resolution and Discussion**

To address algorithmic bias and the proxy problem, simply eliminating sensitive attributes far not reach the root of the issue. One possible key lies in alternating the reasoning process of machines; in other words, breaking down consistent correlations between neutral data representations and the underlying societal structures they reflect (Johnson). The alternation, if successful, can set each attribute separate and restrict algorithmic bias.

However, reducing reliance on proxy attributes risks a tradeoff with judgment accuracy. In some cases, the most rational decisions lie upon the proxy attributes. For instance, Chinese companies recruit mostly based on individuals' educational backgrounds. Yet, urban

residents, who are typically richer and enjoy privileged access to quality education, have better job opportunities (Wu and Treiman). However, if companies eliminate education factors, it would become much harder to accurately identify worthwhile candidates, at least insofar as worthwhile is taken to align with the abilities and qualifications represented by education. At root, what we deem "accurate" is itself filtered through social systems already warped by oppressive structures. Hence, as Dr. Johnson mentions, "[T]here are no purely algorithmic solutions" (Johnson 21). Algorithmic bias cannot be solved in its own system and logic: we have to jump out of the scope.

Turning to algorithmic opacity, the EU's risk-classification law system acts as a great model of how regulations can resolve algorithmic bias - strictly banning and regulating the high-risk AIs, and forcing companies to share more data on the low-risk AIs. The process should not end at increasing transparency; transparency is merely a means towards reaching intelligibility. Only by requiring firms to share understandable judgment processes with the public can give each one of us the path and torch to investigate. The vision lies in a near future in which everyone can have access to algorithms that show their reasoning processes clearly. People treated unfairly by algorithm bias would have a chance to make up for their loss through the legal process. In this sense, an efficient and impartial appeal system is necessary as well; and none of these can be achieved without promulgation of new laws.

**Conclusion**

Out-of-control biases warn us that machines are crossing the boundary between controller and controlled (Fry). But After all, algorithms have been rooted in our society and will inevitably play an increasingly vital role in the world of reputation, research, and finance (Pasquale). Besides, considering the complex tangle between algorithms and human reasoning, and the loss of efficiency and accuracy in judgment processes algorithms are removed, there is no way that we can remove algorithms from our society. The future lies in a world of both humans and machines, where no one can exist without the other.

To balance this vital relationship and regain the controller position, what we need are legal solutions that would increase intelligibility of algorithms.

**Works Cited**


Benjamin, Ruha. *Race after Technology: Abolitionist Tools for the New Jim Code*. Medford, Ma, Polity, 17 June 2019.

Boyd, Danah, and Kate Crawford. "Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon." *Information, Communication & Society*, vol. 15, no. 5, 10 May 2012, pp. 662–679, https://doi.org/10.1080/1369118X.2012.678878.

DLA Piper. "Law in United States - DLA Piper Global Data Protection Laws of the World." *Www.dlapiperdataprotection.com*, 6 Feb. 2025, www.dlapiperdataprotection.com/?t=law&c=US.

Eubanks, Virginia. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor.* New York, St. Martin's Press, 23 Jan. 2018.

European Parliament. "EU AI Act: First Regulation on Artificial Intelligence." *European Parliament*, 19 Feb. 2025, www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence.

Fischer, et al. "Automating Society Report 2020." *Automating Society Report 2020*, 2020, automatingsociety.algorithmwatch.org. Accessed 7 May 2025.

France, Anatole. *The Red Lily*. Charleston, Sc, Bibliobazaar, 2007.

Fry, Hannah. *Hello World : How to Be Human in the Age of the Machine*. London, Black Swan, 2019.

Johnson, Gabbrielle M. "Algorithmic Bias: On the Implicit Biases of Social Technology." *Synthese*, vol. 198, no. 10, 20 June 2020, https://doi.org/10.1007/s11229-020-02696-y.

Office of Privacy and Open Government. "Privacy Laws, Policies and Guidance." *U.S. Department of Commerce*, www.commerce.gov/opog/privacy/privacy-laws-policies-and-guidance. Accessed 3 May 2025.

Pasquale, Frank. *The Black Box Society*. Harvard University Press, 5 Jan. 2015.

Wu, Xiaogang, and Donald J. Treiman. "Inequality and Equality under Chinese Socialism: The Hukou System and Intergenerational Occupational Mobility." *American Journal of Sociology*, vol. 113, no. 2, Sept. 2007, pp. 415–445, https://doi.org/10.1086/518905. Accessed 26 Apr. 2019.