AMPLIFYING BIAS, AUTOMATING RACISM

Ethically framing the issue of algorithmic discrimination

Diletta Goglia

16 July 2020

Abstract. Riding the wave of the current #BlackLivesMatter movement, in this work I present an example of racial-based discrimination resulting from a particular kind of algorithmic bias, known in literature as amplification bias. Then, in order to stimulate debate on how to effectively address this phenomenon in automated decision-making systems, I explore a case study reporting it, I analyse why it is so problematic, and I stress the need to reflect on ethical solutions to fix it.
In this paper, I argue that we cannot solve harmful consequences of algorithmic bias to the root without adopting an ethically-based approach.
Hence my discussion emphasizes the need to urgently address these issues, proposes actions to do that and underlines the importance of rules and norms to keep human beings accountable.

Keywords. Amplification bias, algorithmic bias, AI ethics.

1. Introduction

Machine Learning algorithms are increasingly used in autonomous decision-making tasks in many social domains. [1] However, despite this widespread adoption, still too much optimism goes around these techniques.
It is often stated in writings, how the efficiency of these algorithms will improve performance of many services in different contexts. Unfortunately, it is increasingly common that unwanted consequences of these algorithms arise because of a phenomenon called *bias*.

In its most general sense, the term bias means simply "slant." [2] Given this undifferentiated usage, bias can describe both moral and nonmoral circumstances. My discussion, however, focuses on computer technologies whose bias is a source of moral and ethical concern and, thus, I use the term bias in a more restricted sense. That said, in this work, the term bias refers to a computational product that systematically and unfairly discriminates against certain individuals in favour of others, denying an opportunity or a good on grounds that are unethical.

Starting from this definition, I point out two further considerations as premises for the following argumentations. (1) First, systematic results in algorithms do not establish bias unless they are joined with an unfair outcome. (2) Second, unfair discrimination does not establish bias unless it occurs systematically.

That said, I proceed in analysing an algorithm from a case study, bringing to light a specific kind of bias present in it, known as amplification bias, and I point out how and why it occurs. Then, given the previous premises, I argue the ethical and social fairness of that algorithm and I conclude presenting some solutions to adopt.
It must be said that, to date, there already exist many technical solutions to fix amplification bias: I will mention some of them that, in my opinion, are interesting. However, I will not dwell much on them since my proposal wants to be mostly ethical.

2. Case study

For those not familiar with technical notions as “algorithmic bias” or “objective function”, I will

insert short and simple definitions and explanations for concepts like these.

For now, I will refer to [3] as a starting point for a more general analysis and reflection on amplification bias and, more in general, on algorithmic bias.

This study finds racial bias in an algorithm developed and deployed by UnitedHealth Group that is widely applied by health systems in the care of 70 million patients [4]. It is used by both hospitals and insurance companies and its goal *should be to identify* those patients with the greatest medical need (i.e. high-risk patients who have chronic conditions and "who are at risk of catastrophic complications" [3], that would strongly benefit from additional care).

All this serves to allow the algorithm to make an autonomous decision about whether to include patients in an extra-care programme. [1]

But what the algorithm is actually predicting (what is called the "objective or target function" in Machine Learning) is health *cost*, rather than health need. To better explain: health cost prediction is used as a proxy for health need.

To let you dwell longer on this tricky point, and to give you the time to come up with precious reflections, I will deliberately rewrite the last sentence differently: *developers decided to use* health cost prediction as a proxy for health need.

And this *choice* is exactly the mechanisms by which bias arises: many Blacks patients, whose health is seriously at risk, are excluded.

As confirmation, a fixed unbiased version of the same algorithm suggests that by modifying just the objective function (i.e. not considering expenses *any more* but only medical needs) Black patients included in the extra care programme would increase from 17.7% to 46.5%. [5]

In other words, what is happening, is that healthier Whites are prioritized before sicker Blacks.

It is already evident how much this huge bias is excluding people that could truly benefit from this addictional intervention.

I discuss further below why and how this is happening.

2.1 Problematic assumptions

Reading the study, I noticed that there is an hidden assumption that developers made when they had to decide the objective function to model: it is that if you are spending less money on health care means that you are a healthier person.

Leaving the algorithm aside for a second, I will have no difficulty in convincing you that this assumption is really problematic *per se*. In fact, it automatically excludes a long series of conditions for which people *cannot* spend money on caring for themselves despite being really unhealthy.
I will explore in detail these conditions in 2.2.

The issue arises when hospitals use this algorithm since they are selecting patients likely to cost more in the future, and *not* on the basis of their actual health. Using this "surrogate marker"[2], the algorithm creates a bias against racial and ethnic minorities who don't always get care for the diseases they have. [6]

In other words, instead of being trained to find the sickest, in a physiological sense, this algorithm ended up finding the sickest in the sense of "those whom US society spend the most money on".

[1] Quote from the manufacturer: "the goal of the algorithm is to target patients with complex health needs and to flag them for intervention before their health becomes catastrophic" [3]

[2] Alternative term for "proxy", by Dr. Cardinale Smith, associate professor of medicine at the Icahn School of Medicine at Mount Sinai in New York City.

On this money spent, there exist systemic racial differences. I am going to explore them in more detail in the following paragraph.

2.2 Cost-health disparity

It must be said that, considering cost in isolation (i.e. not considering patient's health), the algorithm has an unbiased and accurate cost prediction, both for Black and White patients. To facilitate this analysis, from now on, I will call this cost-in-isolation simply "cost".

This means that when the algorithm looks ahead to next year's costs and predicts those using data from the current year, it returns correct values.

By the way, this is exactly the double-talk unsheathed by the owner company when the bias-arising issue was brought to their attention: they replied that "the algorithm does what it is designed to do", i.e. predicting cost. And this is actually true.

But it is just as true as it shows race differences in cost-health relationship. In this paper I will refer to this relationship with the term "health cost", in order to distinguish it from the notion of "cost" introduced before.
In fact, despite this fairness in cost, Obermeyer et al. find substantial disparities in health between Black and White patients.

That said, I can safely affirm that, because of the different relationship between cost and health for Black and White patients, by definition, predicting health cost accurately means *not* predicting health accurately, i.e. under-predict health risk for Black patients.

This can sound surprising, because health costs and health needs are highly correlated, *but* it is not always true that there is causality between *needing* health care and *receiving* health care. The even more problematic point is that this disparity is related to race.

As a matter of fact, the result is that Blacks are substantially less healthy than Whites at any level of algorithm predictions, and at the same time and at the same level of health, they result costing 40% less.

The literature broadly suggests two main potential channels from which these disparities in health cost arise.

(1) First, there exist socioeconomic factors correlated with race, like differential insurance coverage, financial barriers and logistical barriers: people with lower incomes typically run up smaller health costs because they are less likely to have insurance coverage, free time, transportation, or job security needed to easily attend medical appointments [3] [4]. Moreover, poor patients face substantial barriers to accessing health care, even when enrolled in insurance plans. [3]
To the extent that race and socioeconomic status are correlated, these factors will differentially affect Black patients.

(2) Second, race can also affect costs directly via several channels: on the way in which black patients are treated by doctors and on racial disparities existing in healthcare, I refer to Section 7.

In the following paragraphs I will address the issue of systemic racism, in particular why it is problematic in an algorithmic perspective.

[3] Actual condition described by Linda Goler Blount, president and CEO of nonprofit the Black Women's Health Imperative.

3. General analysis on algorithmic bias

3.1 Technical analysis

Once I quickly scanned the crucial points of this case study, I want to provide more technical details related to this particular kind of bias that I have just reported.

Literature is crawling with different types of bias that can affect a Machine Learning model.

I state that these categories of bias are not mutually exclusive: any application could suffer from any combination of them. However, identifying and characterizing each one as distinct, makes them easier to tackle. [7]

In the case study examined in 2., I identify two main sources of harm. (1) *Historical bias*, that arises when there is a misalignment between the world *as it is,* and the objectives encoded and propagated in a model. (2) *Measurement bias*, that arises when choosing and measuring features and labels to use; these are often noisy proxies for the desired quantities [7]. "Noisy" in the sense that they may leave out important factors that lead to differential performance.

The most striking proof of how these types of bias are present in the above-mentioned algorithm is the assumption, as well as its consequences, that I have described in 2.1.

The overall resulting bias, that derives directly from the simultaneous operating of these two, is known as *emergent* or *amplification bias*. It is the computational version of confirmation bias[4] and it arises when a computational system reinforces prejudices and "controversial values" [10] that already exist in society, both in an explicit or implicit way.

Amplification bias only arises as context-specific implementation, i.e. in direct application with real users [2]. This makes it even more problematic since it cannot be identified until *after* the deployment of the algorithm in real world contexts.

In particular, the algorithm analysed before has an *implicit* amplification bias since it does not take account of race when predicting health cost [4] , but race factor emerges anyway: as a matter of fact, even if developers had removed race information from data, the algorithm recognizes "Blacks" as feature correlated to patients with lower healthcare costs, and for this reason it disadvantages them.
This phenomenon is called *emergence of correlated feature*. And "the end effect would be almost identical to discrimination through the use of direct race data" [8].

Not only these autonomous decision-making systems can become sensitive to hidden correlations but, according to some studies [9], they seem to *amplify* existing patterns in data and, as a consequence, they also amplify bias contained into it.

3.2 Ethical analysis

At this point, it is reasonable to ask where do these hidden features come from. Well, they come exactly from our society and so they derive directly from us.

I point out that socioeconomic factors and factors directly related to race that I analysed before are all coded in our social behaviour. It is not the algorithm itself that is dangerous or "bad": it just explores data and finds hidden deep beneath it

[4] See also: R. S. Nickerson, *Confirmation Bias*, 1998

the discrimination that we, as human beings, perpetuate long before Machine Learning was born.

Its skewed performance shows how even race-neutral formulas can have discriminatory effects when they lean on data and technical choices that reflect both historical and current inequalities in society. We have seen that not taking care of Black patients in the past, and giving relevance to this situation when choosing the target, led the system to exclude them directly from a healthcare program that would strongly benefit them.

The algorithm *learns* implicit mechanisms of systemic racism and applies them to its computations and predictions, often, as mentioned above, amplifying them. And this is where ethical issues arise: transferring human bias to an autonomous decision of a machine, that not only causes it to emerge but amplifies it, is actually endangering thousands of lives.

Therefore, if we want to address the real cause of this issue going deeply to the root, we find that the danger is in our approach, not in technology *per se*. And, in my opinion, it is exactly in that root that, beyond finding the problem, we find the solution.

I strongly want to stress the fact that the underlying issue relies on the *approach* with which this algorithm was built in order to achieve a certain goal. This is because the goal, coded as objective function, is never given *a priori* but it is always deliberately decided by developers (i.e. by *people*).

Hence, the key point of this discussion is: why was health *cost* chosen? Why developers decided to pursue this kind of approach?

Now I go through the reason why this choice was made, from which social context it derives, and I analyse it as a purely ethical reflection, since this is the main purpose of this paper.

Basically, the principle on which this decision was made is “follow the money”: in fact, this proxy-based approach is typical of the industry-wide strategy.

Unfortunately, the choice of convenient proxies for ground truth is actually proven to be an important source of algorithmic bias in many contexts [3].

Although it is implicitly acknowledged that business model of insurance is “to take more money than you dish out” [5] and so the choice of cost is privately optimal, I claim that from a social and ethical point of view, it is absolutely not.

As a society we care about health, but hospitals, insurers and private companies in general, cynically cares about their own costs.

It is not difficult to realize that *cost* and *need* are not the same thing. Even if in healthcare system they are somehow related, if your goal is to select people based on their health *needs,* you cannot decide to predict health *cost* as your target. More precisely: you can decide it, because it has been done. But is it *ethical*?

Well, I argue that…

It is *not* ethical to pursue profit-making as first purpose in delicate domains like healthcare.

It is *not* ethical to choose proxies as targets for Machine Learning models if it is proven that they lead to discriminations.

It is *not* ethical to deny patients in critical conditions the opportunity to be treated because they are Black.

It is *not* ethical to deploy and continue using autonomous systems in which a racial bias has been detected, especially if their decisions affect thousands of people.

I am aware that this latter claim sounds controversial in Machine Learning community and, more in general, in scientific community: this because, historically, science was conducted based on whether or not we could do something and not based on whether or not it was ethically responsible to do it.
But I strongly support that, if it is proven that an algorithm has harmful consequences on people, its application *must* be suspended. A negative contribution to society can never be translated into a positive contribution to the scientific field.

Given all these argumentations, I conclude that choices made in building the target function of this algorithm are misguided and unethical.

Now someone might reasonably ask if there are any laws or norms to rely on, in order to regulate the behaviour of these private authorities. Well, the answer is no.

Although there exist global associations, non-profit organizations, and important initiatives[5] that are working to build new policies and regulations to face ethical and social issues in AI products, since these are considered "commercial products" there are no appliable laws.

Moreover, lot of norms coming from the legal framework that prevent from discriminating protected classes[6] do not apply for a further reason, that is because these products work within hospital systems, i.e. within private institutions. [3]

This means that, nowadays, if a private actor is using autonomous decision-making algorithms in a socially or ethically not-so-valuable-way it is not possible to appeal to these regulations because, being it private, they do not apply.

In addition, these initiatives themselves still have many issues and open questions: they are mostly principles and guidelines, not real laws. Yet they are works-in-progress, often with internal contradictions, both in the definition of objectives to pursue and in the heterogeneity of internal members. Furthermore, the principles they outline are not always clear, non-ambiguous and simple to follow: their structural complexity does not make them a straight reference to be applied.

4. Solutions

To formalize fairness in a strictly computational perspective there exist many mathematical constraints and statistical methods, like PCA, features decorrelators and adversarial debiasing: in practice, what they do is altering the distribution that the model learns, in classification and decision-making tasks, in order to mitigate bias coming from emergent correlations. Here I mention only one that I find particularly interesting, even if (unfortunately) is still not widely known. It is proposed in [11] and it formalizes explicitly bias amplification phenomenon as the difference between two metrics directly related to protected classes predictions: "model leakage" and "data leakage".

However, my purpose here is to trace out another kind of approach.

The point is not to argue against any particular solution, but rather to frame a possible one in a strictly ethical perspective, just as I framed the problem in the same exact perspective in 3.2.

Although the most common view in Machine Learning community suggests that strictly technical solutions constitute a foolproof system,

[5] Just to mention some of them: FAT-ML, HLEG AI, AAIH, DADM.

[6] The notion of "protected classes" here refers to groups of people with characteristics like sex, race, … who should be legally protected from discrimination.

I argue that an ethical one is more reliable, since it ensures a deeper change to the root. It makes sense to think that science could be a precise and reasonable approach, but we have just seen how fallible human-driven scientific solutions can be (at least without a solid ethical reference behind).

Therefore, referring to the ethical analysis in 3.2, I claim an ethical methodology to approach this issue.

In this last argumentation I will not refer to the question of Machine Ethics or value-sensitive design[7] [12] as well as I will not take a position on that, precisely because I want to focus on providing solutions belonging to the AI Ethics field. Nevertheless, I argue that the distinction between Machine Ethics and AI Ethics is not mutually exclusive: rather I think that the inclusion of both into a unique context or product would be really interesting.

My proposal is exactly the following. Before any other approach, what I strongly recommend is addressing algorithmic bias through precise and appliable laws also, and above all, for private institutions.

I claim to urgently and explicitly define what is acceptable in this context and, especially, what is not.

It is evident that there is a strong need for concrete appliable rules of appropriate conduct so that to legally pursue unacceptable behaviours, coming from *any actor*. I strongly recommend the deployment of precise ethical instructions, that can become a legal standard to be valid in *any context*.

Furthermore, I claim to address within them another fundamental point, that is responsibility allocation. All people involved in developing, selling and deploying algorithms must be held legally accountable for any unfair outcomes and differential treatments if it is proven that these arise from their unethical choices in building the model.

Providing reasons and explanations for a bad output is not enough: since the issue is affecting thousands of lives, who build and deploy these systems must provide adequate and legally accepted justifications for them.

I hope I have shown enough how much these questions are urgent. But in case I had not, I leave you with the open question whether we can still tolerate that thousands of lives are being put at risk to pursue an economic purpose, while who is accountable to this injustice *cannot* be legally addressed.

Since I have widely argued that biased computer systems are instruments of injustice, according to [2], I believe that freedom from bias should be counted among the set of criteria according to which the quality of systems in use in society should be judged. As with other criteria, such as reliability, accuracy, and efficiency, freedom from bias should be held out as an ideal toward which developers must strive.

5. Conclusion

I hope I have clearly shown the urgency to address the issue of harmful consequences of certain forms of bias, as well as the urgency to establish a solid ethical base to address it properly.

This because until we succeed, biased systems are involving the real ethical danger that human lives can be put at serious risk.

I know that it is difficult to respond quickly to this challenge, since no algorithm can ever be totally free of bias. Mentioning [3]: “finding fixes for bias

[7] See also: B. Friedman, *Value-sensitive Design*, 1996

in algorithms — in health care and beyond — is not straightforward".

However, the fact that bias is a necessary and non-avoidable condition in a Machine Learning product must not justify an unfair output as well as the lack of ethical norms to keep human beings accountable.

I hope that my argumentations, and in particular my controversial critique to human choices, will contribute to even deeper future reflections and concrete measures.

Those who develop and deploy algorithms must be able to refer to a specific ethical regulation and must be legally brought into play if responsible for a proven injustice.

Only addressing ethical *choices* of human beings, we can solve the problem at the root.

Who codes matter.

6. References

[1] S. Barocas and A. D. Selbst, "Big Data's Disparate Impact," *California Law Review,* vol. 607, no. 104, 2016.

[2] B. Friedman, "Minimizing Bias in Computer Systems," *SIGCHI Bulletin,* vol. 28, no. 1, pp. 48-51, 1996.

[3] Z. Obermeyer et al., "Dissecting racial bias in algorithm that guides health decisions for millions," *Science,* vol. 366, no. 6464, pp. 447-453, October 2019.

[4] T. Simonite, "A Health Care Algorithm Offered Less Care to Black Patients," *Wired,* October 2019.

[5] M. Evans, "Racial bias found in health care company algorithm," *CBS News,* November 2019.

[6] L. Carrol, "Widely-used healthcare algorithm racially biased," *Reuters,* October 2019.

[7] H. Suresh and J. V. Guttag, "A Framework for Understanding Unintended Consequences of Machine Learning," February 2020.

[8] B. Goodman and S. Flaxman , *AI Magazine,* vol. 38, no. 3, p. 50.

[9] J. Zhao et al., "Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints," 2017.

[10] H. Nissenbaum, "How computer systems embody values," *Computers,* vol. 34, no. 3, 2001.

[11] T. Wang, "Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations," 2019.

[12] R. Dobbe et al., "A Broader View on Bias in Automated Decision-Making," 2018.

[13] T. Zarsky, "The trouble with algorithmic decision," *Science, Technology & Human Values,* vol. 41, no. 1, p. 118–132, 2016.

7. Additional references

K. Fiscella et al., "Addressing Socioeconomic, Racial, and Ethnic Disparities in Health Care", *JAMA,* vol.283, no.19, 2579–2584, 2000.

N. E. Adler and K. Newman, "Socioeconomic Disparities In Health", *Health Aff*airs, vol.21, no.2 60–76, 2002.

K. M. Bridges, "Implicit Bias and Racial Disparities in Health Care", *Human Rights Magazine*, vol. 43, no. 3