The Lab's Quarterly

2020 / a. XXII / n. 2 (aprile-giugno)

CONTATTI
thelabs@sp.unipi.it

I saggi della rivista sono sottoposti a un processo di double blind peer-review.
La rivista adotta i criteri del processo di referaggio approvati dal Coordinamento delle Riviste di Sociologia (CRIS): cris.unipg.it
I componenti del Comitato scientifico sono revisori permanenti della rivista.
Le informazioni per i collaboratori sono disponibili sul sito della rivista:
https://thelabs.sp.unipi.it

ISSN 1724-451X

"The Lab's Quarterly" è una rivista di Scienze Sociali fondata nel 1999 e riconosciuta come rivista scientifica dall'ANVUR per l'Area 14 delle Scienze politiche e Sociali. L'obiettivo della rivista è quello di contribuire al dibattito sociologico nazionale ed internazionale, analizzando i mutamenti della società contemporanea, a partire da un'idea di sociologia aperta, pubblica e democratica. In tal senso, la rivista intende favorire il dialogo con i molteplici campi disciplinari riconducibili alle scienze sociali, promuovendo proposte e special issues, provenienti anche da giovani studiosi, che riguardino riflessioni epistemologiche sullo statuto conoscitivo delle scienze sociali, sulle metodologie di ricerca sociale più avanzate e incoraggiando la pubblicazione di ricerche teoriche sulle trasformazioni sociali contemporanee.

LQ *The Lab's Quarterly*

2020 / a. XXII / n. 2 (aprile-giugno)

MONOGRAFICO

"Il conflitto sociale nell'era dei robots e dell'intelligenza artificiale",
a cura di Mariella Nocenzi (Università degli Studi di Roma "La Sapienza") e Alessandra Sannella (Università degli studi di Cassino e del Lazio Meridionale")

LIBRI IN DISCUSSIONE

MONOGRAFICO

Il conflitto sociale nell'era dei robots e dell'intelligenza artificiale

A cura di

Mariella Nocenzi
(Università degli Studi di Roma "La Sapienza")
Alessandra Sannella
(Università degli studi di Cassino e del Lazio Meridionale")

BIAS AND DIVERSITY IN ARTIFICIAL INTELLIGENCE – THE EUROPEAN APPROACH

The different roots of bias and how diversity can help overcoming it

di *Martina Desole*[*]

Abstract

When we talk about artificial intelligence, it is important to dispel the myths and concerns that humans are creating a new form of intelligence, with its own conscience. Algorithms learn only from the data with which we train them, that's why they resemble very much the structure of thoughts of who will input the training data-sets in the system. This can generate bias. In the context of machine learning, bias can signify that there is a greater level of error for certain demographic categories that received less attentions or about which we have less information or data. AI it's already been used to make decisions on people' life, but currently vast parts of the society are left out from its development which does not capture their experiences or realities. There is a diversity crisis in the AI sector including gender, visible minorities, race, persons with disabilities, and age. This leads to a problem of inclusion and equity as well, with many people being potentially excluded and disempowered by the creation of probable bias in the technology. The European Commission addresses the issue of equity, diversity and inclusion in the White paper on Artificial Intelligence published the 18th of February 2020, giving a policy framework to implement actions in this direction.

Keywords

Artificial intelligence; Bias; Diversity; Human Centric AI

[*] MARTINA DESOLE lavora all'Agenzia per la Promozione della Ricerca Europea (APRE) Head of International Cooperation National Contact Point H2020 - Nanotechnologies, Advanced Materials, Biotechnologies and Productions.

Email: desole@apre.it

https://doi.org/10.13131/1724-451x.labsquarterly.axxii.n2.129-142

1. INTRODUCTION

In the context of machine learning, bias means that a system's predictions introduce a greater level of error for certain demographic categories that received less attentions or about which we have less information or data. There is not one single root cause for bias, and there are numerous variables that researchers must consider when developing and training machine-learning models and building training data sets. Data sets used by AI systems (both for training and operation) may suffer from the inclusion of inadvertent historic bias, incompleteness and bad governance models. The perpetration of such biases could lead to unintended (in)direct prejudice and discrimination against certain groups or people, potentially exacerbating prejudice and marginalization.

Diversity in development teams can tackle some of the direct and indirect bias. The European Commission, coherently with its Human-centric AI approach, and with the latest Communications and Guidelines on ethical AI, in the recently published White Paper states: "requirements to take reasonable measures aimed at ensuring that [the] use of AI systems does not lead to outcomes entailing prohibited discrimination." This document paves the way for the future regulatory framework for AI, that will define the mandatory legal requirements to be imposed on the relevant actors, and already calls on the request of AI to work as a white box, where the decision-making process of the algorithm can be retrieved at any time. This will also help identifying when bias did occur and help rectify the process. This article has been compiled based on the analysis of available articles and other sources which are addressing the issue together with personal view.

2. EXAMPLES OF BIAS AND DISCRIMINATION IN AI

When we talk about artificial intelligence, despite the name, it is important to dispel the myths and concerns that humans are creating a new form of intelligence, with its own conscience.

Algorithms learn only from the data with which we train them, that's why they resemble very much the structure of thoughts of who will input the training data-sets in the system. This can generate bias. The first example of a chatbot gone rogue, is the story of Tay. Released on Twitter in March 2016 under the handle @TayandYou, Tay, an acronym for 'thinking about you', was built to mimic the language of an average American teenager girl. Capable of interacting in real time with Twitter users, Tay was learning from its conversations to get smarter over time.

Twitter users realised Tay had 'repeat after me' feature enabled. This simple form of machine learning allowed Tay to learn from its interactions and, once targeted with racist and hate tweets, it went from the giggly "Humans are super cool!" to becoming a racist nazi, with her last tweet "Hitler was right." It was actually built following the structure and the example of Xiaoice, the Chinese chatbot that never changed its behavior as the tweet she received were controlled and censored.

This episode is an exemplary case that shows that AI is defined by its training, and training is nothing but a polarization towards its objectives. But if the polarization is towards incomplete (or reprehensible) objective it will produce a bias. We can paraphrase the saying "garbage in – garbage out" with "bias in – more bias out".

There are many examples of bias and discrimination occurred in the past few years. For example, a study by the University of Virginia examined the trend of photographic recognition software to associate images of people in the kitchen with the female sex. Researchers found that Microsoft and IBM's facial recognition services were more accurate with white people than African-Americans. One of the most famous cases involved Google Photos: the company had to completely revise the algorithm that associated images of African Americans with the label "gorilla". A study cited by the Guardian showed that the same CV, analysed by the AI, was 50% more likely to get a job interview when the candidate had a "Euro-American" name compared to an "African-American" one. On November 11, Apple was accused of developing an AI software for credit lines which discriminated against the wife compared to her husband, guaranteeing her a lower credit even if they had the same assets. And the examples can continue targeting bias in gender, race, status, and other less represented demographic categories.

3. UNDERSTANDING THE BIAS

In the context of machine learning, bias in the results can mean that there's a greater level of error for certain demographic categories. The causes of this type of bias are not univocal, there are numerous variables that researchers must take into account when developing and training machine-learning models in order to avoid bias in the results.

3.1. *Incompleteness of the input data*

To understand the phenomenon of bias it is important to understand that the vast majority of Artificial Intelligence systems are based on a

machine learning algorithm whose fundamental goal is to provide a classification of the input data. It might be a simple image recognition algorithm or a complex financial analysis tool, but in the end it will try to guess if the input data belongs to one category or another. In the first example it will try to estimate to which of the known categories the image is more similar, in the second it will try to estimate if the behaviour of some stocks belong to the condition where it is better to sell or to buy. If the training has been affected by bias, the categories known by the systems will be incomplete and when the algorithm will try to make a guess on data whose category had not been part of its training it will be a wrong one. To make an example we can imagine a system that has been trained to decide if an image is a cat or a dog. This has been done training a neural network with thousands of pictures of cats and dogs and labelling each picture with the correct corresponding category.

The system will work wonderfully with pictures of cats and dogs, but if we input the image of a rabbit the algorithm (who has never seen a rabbit) will have no valid option and will randomly guess that it's a weird cat or a fluffy dog. This type of incorrect behaviour is representative of a bias due to incompleteness of the training dataset. Origins of bias can always be found in an incomplete training dataset. When the dataset does not contain all demographic categories it will work fine with data belonging to the limited categories known but it won't scale properly when the variance of data is increased: a system trained to recognize only white people will have issues recognizing Afro-American people; a system trained to understand vocal commands from male users will have a higher rate of error with female users; similarly, a system capable of understanding voice commands from English native speakers will be unsuitable to be extended to foreign speakers.

Dangerous biases due to incomplete datasets might affect specific components of a society that were not considered when the model was trained. This type of bias can be avoided if the dataset is prepared with accurate statistical knowledge of the population and the corresponding classification categories; when the use of a particular model is extended to a wider or different population, retraining of the model has to be considered. These types of models might be applicable for the original scope but don't scale properly when applied to a larger population.

3.2. *Bias in the data labelling*

Another possible cause of bias does not come from an incompleteness of the data, but from a bias in the classification. We can imagine a

system trained to distinguish pests from innocuous insects on crops. If the operator has a fear for spiders it will label all spiders as pests and maybe butterflies as innocuous. This will result in a system that will try to exterminate good spiders and will spare dangerous butterflies. Something worse might come in case of a prejudice embedded in the labels where a prejudiced category would suffer the prejudice encoded in a black box model which does not and cannot give explanations of its choices. This is one of the stronger reasons that advocates for an Explainable Artificial Intelligence (XAI) because, especially in the field of AI justice and AI administration, everybody has the right of an explanation. XAI has the goal of developing Artificial Intelligence systems that overcome the limitations of black boxes and where the training and the logic behind every decision is transparent to the users.

3.3. *Incompleteness of the model*

Finally, another possible source of bias might come from the choice of the algorithm itself. Not all Artificial Intelligence models are the same, and the selection of one over another might result in different outputs. This risk is mitigated with an extensive and scientific approach to test and validation of the models in order to asses that the chosen model correctly maps the desired categories and that there is no unexpected input capable of producing an mistaken result. One possible cause of bias in Artificial Intelligence systems comes from an incompleteness of the model. When doing a piece of scientific work the data collection process is a critical part of it. But when we create a model of some kind, the data are only an input to it. So, we need to verify if the model is accurate before any of its assumptions can be verified. Another cause of bias comes from the test used. An AI model is only as good as its results, therefore it is only as good as its test set. To improve it, we could use more data or more accurate tests. Most of the time, however, even with all the more perfect testing and new data, we can't achieve even a basic competence of AI.

4. THE "WHITE GUY PROBLEM"

Beside the introduction of more accurate tests, another reason behind bias in AI it is also deemed to the composition of the development team.

The lack of diversity can bring inherent bias in the composition of the datasets or the labelling: this is called the "White guy" problem. The

typical AI developer is male and white. This is not a reflection of the native population of developers but rather the members of the first generation of AI developers who come out of the system engineering field and not with different technical backgrounds.

A recent report from the AI Now Institute (2019) found that 80% of AI professors, 85% of AI research staff at Facebook, and 90% of those staffers at Google are male. Further, people of color make up only a small fraction of staff at major tech companies. There is a gender gap in the participation of women in all the STEM fields, in Information Technologies, women make up only 24% of the users of coding platforms (or about 20% of the total number of active users) (cfr. OECD 2018). With current rates of coding being below the average of men, that still amounts to 40% of active users being female. The percentage drops when it comes to women in machine learning, that accounts only to 12% of all developers. This shortfall in diversity can lead directly to shortcomings in the resulting technology. The AI industry is also recognizing that bias would hinder the capacity of performing solid and effective predictions if the diversity issue is not tackled properly, by encouraging mistrust and producing distorted results. Therefore, there is a need for a profound shift than can start from ensuring more diversity in the developer teams. To date there is still a lack of data to be able to analyse the other minorities involved in AI.

Graf. 1. Gender Balance in machine learning

Source: Element AI – Global AI Talent Report 2019

4.1. *Developers Team diversity*

More diversity in teams can help tackling the inherent direct and indirect biases, because they will realize when datasets and models skew toward inadvertent bias. Diversity means building inclusive teams with diverse backgrounds to integrate unique perspectives, and considering that a successful team is also the result of diverse competencies, including social scientist.

Measures suggested to tackle the issue envisage, among others, to encourage machine-learning teams to measure accuracy levels separately for different demographic categories and to identify when one category is being treated unfavorably. These actions resulted easier when diverse teams have been performing this "de-biasing" action. It is not only the small percentage of women in developers' teams, but the overall lack of diversity of visible minorities, race, ethnicity, persons with disabilities, and age.

As also stated in the Science for policy report by the JRC, to build trust in the AI applications, it is needed to have a responsible approach in AI design, starting from the team composition until the diversity in the training dataset (Craglia 2018: 61). Despite the fact that AI it is already been used to make decisions of the life of people, currently vast parts of the society are left out from its development, which when implemented does not capture their experiences or realities. This leads to a problem of inclusion as well, with many people being potentially excluded and disempowered by the creation of probable bias in the technology as well.

5. EQUALITY, DIVERSITY AND INCLUSION: HOW TO MINIMIZE THE SOCIAL IMPACT OF THE BIAS IN AI

In the previous chapters, we have analysed examples and possible causes of bias in AI that can lead to discrimination for gender, visible minorities, race, persons with disabilities, and age. These causes can be synthetized in two main categories: biased data sets and lack of diversity in the teams. Beside the possibility to use debiasing methodologies for the data set, a real diversity in teams of developers can be consider a solution for tackling the inherent bias in less inclusive teams. We can come to the conclusion that underrepresented categories in AI developer teams might suffer from discrimination, and this is a matter of equality and inclusion. As AI is now used in many crucial domains, discrimination could have an impact on getting hired, not receiving parole, having different rates for

a loan or an insurance, or having a wrong medical diagnosis. The new approach to help minimizing the bias, and therefore the potential social impact of an unfair AI is taking into consideration best practices in equality (and equity), diversity and inclusion (EDI) at every development stage, from research design, to team composition, to present outcomes, and dissemination. The EDI approach, together with the "ethic" by design proposed by the European Commission (see below), are by now pursued in many research Institutes, for example in Europe in the University of Bristol, where the curriculum of the Centre for Doctoral Training for Interactive AI includes "material on equality, diversity and inclusion in the design and use of AI: for example, potential issues with demographic bias in AI algorithms, and mitigations", and it is shared also with their industrial partners. Or in the Alan Turing Institute[1] that established an EDI Advisory Group «to ensure that Turing is effectively addressing equality issues and complying with relevant legislation by giving strategic direction and overseeing the continuing application and development of EDI policies in line with legislation and strategic objectives». This approach has been pioneered across the Atlantic though, being Canada the first Country to launch a policy of ethics in AI, with the adoption of the Montreal Declaration on Responsible AI, launched the 28th of February 2018. IVADO[2] the Canadian Institute for Data Valorisation champions equality, diversity and inclusion both in its own institute and through the promotion of these principles to its ecosystem of researchers and industrialists, has been among the advocates for the creation of the *Observatoire international sur les impacts sociétaux de l'IA et du numérique* (OBVIA) to measure the impacts of AI helping communities, organizations and individuals to maximize the positive spinoffs of AI and digital technology and to minimize the negative effects of technologies.[3] It is important to consider that as a biased data set produced by a non-diverse developers team can lead to bias, we can actually use a well-trained AI to fight inequalities and discrimination. The EDI approach is helping training the new generation of researchers that will be hired by multinationals, industries or start-ups, getting there with a set of skills that will allow to avoid to perpetrate

[1] The Alan Turing Institute is the national institute for data science and artificial intelligence, with headquarters at the British Library: https://www.turing.ac.uk/about-us/equality-diversity-and-inclusion.

[2] Canadian Institute for Data Valorization: https://ivado.ca/en/about-us/equity-diversity-and-inclusion/.

[3] International observatory on the societal impacts of AI and Digital: https://observatoire-ia.ulaval.ca/.

human bias. But, taking it even further, AI can become an active force for good and being used to defy inequities and prejudices. For instance, an artificial intelligence tool for detecting unfair discrimination – such as on the basis of race or gender – has been created by researchers at Penn State and Columbia University in 2019 and many others followed, like Aequitas, an open source toolkit aiming to track and correct biases in databases and machine learning models, developed by the Center for Data Science and Public Policy of the University of Chicago.

6. THE EUROPEAN HUMAN-CENTRIC AI APPROACH

In June 2018, the European Commission sets up an indipendent High Level Experts Group on Artificial Intelligence (HLGAI), who in 2019 published the Ethical guidelines for Trustworthy AI. In the spirit of the open and transparent co-creative European approach to policies, a first draft of the document was released on 18 December 2018 and was subject to an open consultation which generated feedback from more than 500 contributors. Europe has also taken on a coordinating role of the Member States, asking each State to develop its own strategy, which conveyed into the recently drafted White Paper on Artificial Intelligence published on the 19th of February 2020 (EC 2020: 19). In general, Europe has taken on a coordinating role to ensure that the approach of the various European countries is also based on European values. This approach on AI and robotics aims to deal fully with technological, ethical, legal and socio-economic issues to support European industrial and research capacities and to ensure that AI is at the service of citizens and the economy European. The Commission therefore launched the concept of "Human centric AI", that places people at the center of the development of AI (Commission to the European Parliament, The Council, The European Economic and Social Committee and the Committee of the Regions Brussels 2019), focusing heavily on the future impact of technologies and especially on three fundamental actions:

- support and promote European capacities for development and adoption of AI technologies;
- prepare and prepare for socio-economic change;
- ensure an adequate legal ethical framework;

Within the European guidelines, fundamental themes are certainly:

- socio-economic impact (how society will change and how the world of work will change and we will deepen this point shortly);

- openness and democratization (in the sense that AI must be made available as much as possible through platforms where knowledge can be shared openly);

- ethics by design – the consideration of ethical principles in all stages of technology development, with the aim of building a relationship of trust with civil society and ensure that the slogan "AI for good and for all" becomes a reality.

6.1. *Diversity in the Seven Principles*

In Ethics Guidelines for Trustworthy AI, the HLGAI postulate that in order to achieve "trustworthy AI", three components are necessary: (1) it should comply with the law, (2) it should fulfil ethical principles and (3) it should be robust. Based on these three components to ensure that European values are at the heart of creating the right environment of trust for the successful development and use of AI Human agency and oversight, the Guidelines identify Seven principles:

1. Human agency and oversight
2. Technical robustness and safety
3. Privacy and data governance
4. Transparency
5. Diversity, non-discrimination and fairness
6. Societal and environmental well-being
7. Accountability

The principle of Diversity, non-discrimination and fairness includes also the avoidance of unfair bias, accessibility and universal design, and stakeholder participation.

The HLGAI, states that "in order to achieve Trustworthy AI, we must enable inclusion and diversity throughout the entire AI system's life cycle. Besides the consideration and involvement of all affected stakeholders throughout the process, this also entails ensuring equal access through inclusive design processes as well as equal treatment. This requirement is closely linked with the principle of fairness".

The Guidelines sets out a non-exhaustive Trustworthy AI assessment list (pilot version) to operationalise Trustworthy AI addressing developers and deployers of AI, to help assessing the trustworthiness of their processes/results. It also provides general recommendations on how to implement the assessment list for Trustworthy AI though a governance structure both at operational and management level. This should point all the users in the right direction to comply with an ethical approach to AI.

5. Diversity, non-discrimination and fairness

Unfair bias avoidance:
Did you establish a strategy or a set of procedures to avoid creating or reinforcing unfair bias in the AI system, both regarding the use of input data as well as for the algorithm design?
♣ Did you assess and acknowledge the possible limitations stemming from the composition of the used data sets? ♣ Did you consider diversity and representativeness of users in the data? Did you test for specific populations or problematic use cases? ♣ Did you research and use available technical tools to improve your understanding of the data, model and performance?♣ Did you put in place processes to test and monitor for potential biases during the development, deployment and use phase of the system?

Depending on the use case, did you ensure a mechanism that allows others to flag issues related to bias, discrimination or poor performance of the AI system?
♣ Did you establish clear steps and ways of communicating on how and to whom such issues can be raised? ♣ Did you consider others, potentially indirectly affected by the AI system, in addition to the (end)- users?

Did you assess whether there is any possible decision variability that can occur under the same conditions? ♣ If so, did you consider what the possible causes of this could be? ♣ In case of variability, did you establish a measurement or assessment mechanism of the potential impact of such variability on fundamental rights?

Did you ensure an adequate working definition of "fairness" that you apply in designing AI systems? ♣ Is your definition commonly used? Did you consider other definitions before choosing this one? ♣ Did you ensure a quantitative analysis or metrics to measure and test the applied definition of fairness? ♣ Did you establish mechanisms to ensure fairness in your AI systems? ♣ Did you consider other potential mechanisms?

This also means to avoid, if possible, the use of "black-box models" (EC 2020: 13) and ensure transparency both with a "white box" model approach (where the decision making process of the algorithm can be retrieved at any time) and the by «keeping of records and data documentation on the programming and training methodologies, processes and techniques used to build, test and validate the AI systems, including where relevant in respect of safety and avoiding bias that could lead to prohibited discrimination» (ivi: 19).

The European AI strategy make clear that trust is a prerequisite to ensure a human-centric approach to AI: considering "AI not an end in itself, but a tool that has to serve people with the ultimate aim of increasing human well-being. To achieve this, the trustworthiness of AI should be ensured. The values on which our societies are based need to

be fully integrated in the way AI develops» (High Level Expert Group on Artificial Intelligence 2019: 1). To achieve this, an ethical by design approach should be ensured at every development stage and as stated in the "diversity in terms of gender, racial or ethnic origin, religion or belief, disability and age should be ensured at every stage of AI development" and the European Commission is going to enforce a strong regulatory framework that will set the global standard for humancentric AI.

CONCLUSIONS

The reasons for bias in AI are diverse and the previous sections show some of the most common reasons for bias. The lack of representation of minorities, including gender, race, class, disability, age in development teams is one of the outstanding issues that should be tackled immediately. Algorithms learn only from the data we train them with, that's why they resemble, very much the structure of thoughts of who will input the training data in the system. Among different reasons like incompleteness and classification of training data, we must not forget that a lack of representation of minorities among programmers/decision makers/investors can generate also problems of accessibility and inclusion. We have an obligation to create technology that is effective, accessible and fair for everyone. In order for the benefits to outweigh the risks, it is important that policy makers, industry leaders, researchers, strive to maintain high attention, to seek and develop solutions that reduce prejudice towards all minorities. The European Commission is calling for an Ethical Human-centric AI, where the issue of bias and diversity are clearly addressed and the first steps towards a common regulatory framework are set in place.

The consideration of ethical principles in all stages of technology development secures a relationship of trust with civil society and ensure that the slogan "AI for good and for all" becomes a reality.

BIBLIOGRAPHICAL REFERENCES

AI NOW INSTITUTE (2019). *Discriminating Systems Gender, Race, And Power In Ai*. New York: New York University.

COMMISSION TO THE EUROPEAN PARLIAMENT, THE COUNCIL, THE EUROPEAN ECONOMIC AND SOCIAL COMMITTEE & THE COMMITTEE OF THE REGIONS BRUSSELS, 8.4.2019 COM (2019) 168 final, *Communication: Building Trust in Human-Centric Artificial*

Intelligence, https://ec.europa.eu/digital-single-market/en/news/communication-building-trust-human-centric-artificial-intelligence.

CRAGLIA M. (2018, ed.). *Artificial Intelligence An European perspective*. Brussels: European Commission "Joint Research Centre".

EUROPEAN COMMISSION (2020). *A European approach to excellence and trust. White paper On Artificial Intelligence*. Brussels, 19.2.2020: https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf.

HIGH LEVEL EXPERT GROUP ON ARTIFICIAL INTELLIGENCE (2019). Ethics guidelines for trustworthy AI: https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines.

OECD (2018). Bridging the digital gender divide. In Ai Now Institute, *Discriminating Systems* Gender, *Race, and Power in AI*, cit., 2019.

PENN STATE & COLUMBIA UNIVERSITY (2019). Using artificial intelligence to detect discrimination. *ScienceDaily*. 10 July: www.sciencedaily.com/releases/2019/07/190710121649.htm

Numero chiuso il 30 giugno 2020

ULTIMI NUMERI

2020/XXII(1) (gennaio-marzo)

2020/XXII(2) (aprile-giugno)