BIAS, TRANSPARENCY AND EXPLICABILITY.

Carlos Arbonés
B.S. Data Science and Engineering
Universitat Politècnica de Catalunya
carlos.arbones@estudiantat.upc.edu

Núria Arqués
B.S. Data Science and Engineering
Universitat Politècnica de Catalunya
nuria.arques@estudiantat.upc.edu

Marc Colomer
B.S. Data Science and Engineering
Universitat Politècnica de Catalunya
marc.colomer@estudiantat.upc.edu

Pere Cornellà
B.S. Data Science and Engineering
Universitat Politècnica de Catalunya
pere.cornella@estudiantat.upc.edu

Sílvia Fàbregas
B.S. Data Science and Engineering
Universitat Politècnica de Catalunya
silvia.fabregas@estudiantat.upc.edu

Jordi Farràs
B.S. Data Science and Engineering
Universitat Politècnica de Catalunya
jordi.farras@estudiantat.upc.edu

Mauro Filomeno
B.S. Data Science and Engineering
Universitat Politècnica de Catalunya
mauro.filomeno@estudiantat.upc.edu

Marc Franquesa
B.S. Data Science and Engineering
Universitat Politècnica de Catalunya
marc.franquesa@estudiantat.upc.edu

Ignacio Gris
B.S. Data Science and Engineering
Universitat Politècnica de Catalunya
ignacio.gris@estudiantat.upc.edu

Maria Sans
B.S. in Philosophy
Universitat de Barcelona
msansban7@alumnes.ub.edu

Kiranjitsingh J Dogra
B.S. in Philosophy
Universitat de Barcelona
kdografe81@alumnes.ub.edu

Joan Escat
B.S. in Philosophy
Universitat de Barcelona
jescatna7@alumnes.ub.edu

Ana Beatriz de Queirós
B.S. in Philosophy
Universitat de Barcelona
adequeno28@alumnes.ub.edu

***Abstract*—In our contemporary era, decision-making processes increasingly rely on algorithms and artificial intelligence (AI). However, the perpetuation of biases within these algorithms can lead to discriminatory or unjust outcomes, amplifying disparities that often remain unknown to their developers. This article examines the ethical implications of biased AI systems, utilizing real-world cases and debates within a university classroom, involving 50 students enrolled in the Degree in Data Science and Engineering at the Universitat Politècnica de Catalunya (UPC). It underscores the critical role of explainability, transparency, and awareness in addressing and mitigating bias. By identifying inherent biases in algorithmic decision-making and advocating for proactive measures like sensitivity analysis or integrating fairness criteria, the study highlights the need for greater accountability in AI design and implementation.**

***Index Terms*—Artificial Intelligence, Bias, Transparency, Explainability, Ethics, Healthcare, Finance, Recommender Systems**

I. Introduction

Artificial intelligence has advanced rapidly during the past decade, permeating many aspects of our daily lives. From smart homes and IoT to social media, entertainment, and the emergence of large language models like ChatGPT; AI is becoming integral to daily life. However, we have been surrounded by algorithms for quite a long time now, yet we are largely unaware of how the majority of algorithms around us operate and affect us. This phenomenon contributes to the notion of "The Black Box Society", a concept drawn from Frank Pasquale's book ("The Black Box Society: The Secret Algorithms that Control Money and Information" [1])

A "Black Box Society" is one where governments and institutions closely monitor people, and the populace remains unaware of how much information is collected, how it is used, and the potential consequences of this data collection and usage. In the era of digital surveillance capitalism, companies have unprecedented access to personal data through web interactions and social networks. Data brokers and aggregators buy, sell, and trade large quantities of personal data, often without the users' knowledge or consent, and not always with the best intentions.

This combination of lack of transparency and awareness difficulties the detection and correction of biases implicit to the model: out of sight, out of mind. The pervasive use of algorithms in critical areas such as finance, medicine, or social media amplifies these risks, leading to consequences such as the perpetuation of inequalities and new channels of discrimination against individuals, all while remaining invisible to the affected ones.

This article

- Explores the analysis of bias through its categorization in data, algorithms, and socio-technical factors, and examines its implications on structural and social issues from a perspective of vulnerability.
- Aims to raise awareness about the impact of malprac-

tice, which can be avoided through the assumption of responsibility and accountability.
- Highlight the lack of awareness regarding the algorithms that surround us and the biases they include, their functioning, and their impact, even among students of the Data Science and Engineering degree.
- Argue that it is primarily through explainability and transparency that biases can be detected and corrected, and study the possible limitations of these approaches.

To conduct the evaluations and reflections mentioned earlier, a small study was performed in the context of the Ethics class for the Data Science and Engineering degree at UPC. Various materials were presented to 50 students, including three cases that explore different perspectives and issues.

To conduct the evaluations and reflections mentioned earlier, a small study was performed in the context of the Ethics class for the Data Science and Engineering degree at UPC. Various materials were presented to 50 students, including three cases that explore different perspectives and issues.

These cases are later discussed in this article and delve into the topics of:

- Medicine and data bias
- Finance and loans
- Recommendation systems and political elections

These combined make it possible to propose a more general case that will be analyzed by students in connection with several points adjacent to the topic of bias.

II. Type of bias in AI systems

A bias is a disproportionate weighting in favor of an idea or thing, which usually leads to something inaccurate, discriminatory, prejudicial, unfair or outright false, and it is behind every conscious and unconscious decision taken. Biases in AI can be broadly classified into three main types: data bias, design bias, and socio technical factors. [2] Each of these categories plays a role in shaping the behavior and decisions of AI systems. This categorization allows for a detection of biases introduced into a system and work towards mitigating them.

- **Data Bias:** When the data used to train an AI system does not adequately represent the entire population (lack of data or lack of explainable variables), the algorithm's decisions can favor the group that is overrepresented in the training data. These biases can reinforce discrimination and prejudice towards minorities. More in particular, on the internet, a very small subset of users and sources generate most of the content on the web. For instance, 4% of users in Amazon produce 50% of the reviews on the platform [3]. Sometimes, the data bias might represent a natural bias or latent bias. If a disease is predominantly diagnosed in women and a representative sample is obtained, the model may be biased towards the female profile, potentially making errors for other groups.
- **Prejudice in Design:** Implicit biases held by AI designers can unintentionally influence the behavior of the system. This includes algorithmic biases that are embedded in the design and development process. For instance, in 2014, Amazon began the development of an automated system aimed at rating job applicants on a scale of one to five stars. However, the project was abandoned when it became apparent that the system exhibited a bias towards favoring male candidates for technical positions and the system was unintentionally trained to favor male applicants over female candidates [4].
- **Socio-technical Factors:** These biases stem from the influence of social, economic, and cultural contexts on the design, deployment, and use of AI systems. They are the most numerous because they emerge when the model interacts with the real world.

III. Ethical dilemmas in biased AI systems

The integration of biased AI systems into various sectors poses significant ethical challenges. The following points provide a framework for gaining a better understanding of this issue from multiple perspectives.

A. Socio-Economic Issues and Vulnerability

Algorithms used for decision-making can exhibit biases, especially when relying on demographic variables like age and income, which often favor wealthier areas and neglect less affluent ones. Even without explicit demographic data, implicit biases can still arise from prioritizing certain factors, potentially excluding marginalized communities.

Implicit bias involves unconscious negative attitudes toward certain social groups, shaped by learned associations with traits such as race or gender. These biases influence behavior unconsciously and often predict it more accurately than conscious beliefs. This systematic bias is embedded in our reality, influencing algorithms to make discriminatory decisions without developers' awareness. These biased decisions can inadvertently reinforce existing disparities because they align with normalized views and behaviors, making them difficult to detect and correct.

As algorithms are increasingly used for important decisions, their biases can have significant impacts, leading to unequal treatment, limited opportunities, and rights violations. Therefore, we must recognize that nearly all algorithms are susceptible to bias, reflecting our own prejudices. [5] Addressing these biases is essential to minimize their harmful effects, even though we can't eliminate them entirely.

B. Post-conventional ethics

When confronted with inevitably biased algorithms, one must make a decision and face the ethical dilemma of whether the algorithm should be used or not. Even if an algorithm is highly beneficial for a business or a large segment of the population, its bias may have adverse

effects on others. Balancing these benefits and harms is a critical decision that requires careful consideration.

Unfortunately, although there are no universal ethical principles to rely on, one could consider the perspective of Lawrence Kohlberg. In his psychological theory of moral development, he asserts that at the most mature, evolved level, individuals prioritize universal human ethics and the social contract. This means that in post-conventional moral reasoning, the life and dignity of individuals, as well as the well-being of society, are paramount. [6]

Currently, the use of biased algorithms presents a significant challenge in the context of AI, as it can perpetuate and even exacerbate existing inequalities. Ethical decision-making in AI requires a personal and global commitment to mitigating these biases and ensuring fairness and justice. It involves questioning whether the benefits to some justify the harms to others and seeking ways to minimize negative impacts. The future requires post-conventional ethics.

C. Responsibility, culpability, and their impact.

Ignorance regarding the malfunctioning of an algorithm does not absolve us from culpability. In the development of an algorithm, it is imperative to comprehend all the intricacies that may culminate in an adverse outcome. The notion of strategic ignorance refers to those practices that intentionally endeavor to reject pertinent knowledge, thereby enabling a denial of responsibility in the event of a catastrophe [7]. In any case, the responsibility does not fall on those who are knowledgeable. Nonetheless, a formal definition of responsibility and blame is first necessary. Socially, peoples' responsibilities are those things for which they are accountable [8], while culpability is the moral fault accompanying a tortious act that renders the behavior criminal [8]. In the process of crafting an algorithm, individuals must remain conscious of their profound responsibility. A diligent commitment is vital to ascertain the absence of any malfunctioning. However, the delegation of responsibilities is far more intricate than it appears. For example, to what degree are we accountable for the biases embedded within the data? Individuals will invariably be accountable for their actions, whether beneficial or detrimental. However, it is imperative to continually strive to identify potential issues and address them in a manner we deem ethically sound, ensuring we do not bear moral culpability.

D. Purpose: A Goal Chosen Through Will

Technology is always created with a specific goal or purpose in mind, reflecting the values and intentions of its creators. This means that technology is not neutral but is influenced by the ethical and moral principles of those who design and implement it. Every technological tool or system is developed to achieve a particular outcome, benefiting some while potentially disadvantaging others. The resources, education, and support available to the creators play a significant role in shaping the technology, as do the ways in which users interact with it. These interactions can introduce biases, whether intentional or unintentional, into the technology. Consequently, every technological innovation carries with it the potential for both positive and negative impacts, shaped by the values and goals embedded in its design and use.

E. Technology is not axiologically neutral

Technology is not axiologically neutral, for it always serves a human purpose. The term "axiological" encompasses everything related to values, principles, and value judgments, thus the previous statement asserts that human values, in terms of beliefs or principles, are reflected in the conceived technology. At the same time, behind every action there is someone who benefits from it and someone who may be harmed by it; there is someone who was able to implement the technology (be it because they had the resources, education, or external support); there is someone who uses it, who interacts with it (i.e. the web). It is in the act of creation and the subsequent interaction that biases are introduced; whether through the tool's purpose and conceiving, its use, or misuse.

F. Challenges in AI Transparency and Explainability

In the context of machine learning, transparency refers to the clear, understandable, and justifiable explanation of a model's training and prediction processes [9]. Explainability, on the other hand, pertains to providing humans with the understanding and intellectual oversight of these processes [10]. Both of these, despite proving essential to combating biases, do provide certain limitations. Firstly, transparency reveals the inner workings of models, which adversarial parties can exploit. Additionally, the pursuit of transparency and explainability often makes the process much more complex [11]. Furthermore, while the end goal is to foster trust, understanding the model's processes does not necessarily imply trust.

Despite these challenges, transparency and explainability are essential in certain situations. In the context of AI safety, not understanding a model can easily lead it to cheat when solving actual tasks (some examples: [12]). Balancing the benefits with the added complexity is vital for responsible machine learning development.

IV. Case studies

This section delves into real-world cases spanning diverse topics to illustrate the pervasive presence of bias and the lack of transparency or explainability in various contexts. Through examination, the analysis aims to reveal these issues' manifestations and implications, highlighting the need for greater awareness and accountability in addressing bias and transparency. These cases will then serve as a basis for a practical class, where their effects will be debated from multiple perspectives, fostering a comprehensive understanding of the arguments and viewpoints of all parties involved.

A. Medicine

The exponential growth of data and the fast progression of technology have prompted various sectors to integrate artificial intelligence into their operations. The medical field is no exception. To enhance efficiency and increase incomes, medical services have begun automating certain decisions through the use of sophisticated decision models.

In the medical domain, these decision models utilize specific data inputs and process them to produce predictions that ideally align with the decisions a doctor would make. This integration aims to streamline medical processes, ensuring that automated decisions support and mirror the expertise of healthcare professionals [13] [14] [15].

These decision models are already in use in the United States - among several other countries around the globe such as Xina, Italy, Spain or Germany - and several issues have emerged from their implementation. In particular, a risk prediction algorithm designed to help hospitals and insurance companies identify patients who could benefit from a high-risk care management program has exhibited racial bias [16]. This bias arose because the algorithm relied on a metric that used a flawed variable to determine the need for treatment, highlighting significant concerns about fairness and accuracy in healthcare decision-making.

While these algorithms may indeed enhance efficiency, they also risk transgressing ethical standards by perpetuating socio-technical biases inherent within the data. This highlights the task of managing such biases rather than eradicating them, particularly concerning principles of justice and equity. Moreover, responsibility is paramount. Unawareness of a model's intricacies does not absolve us of responsibility for any resulting malfunctions or unethical conduct. Some argue that transparency and explainability offer solutions to these issues, and while this holds partial truth, achieving complete transparency and interpretability sparks debates that delve into the realm of privacy [17]. This not only relates to the field of discoveries but also impacts aspects such as rankings and evaluations. The revelation of rules may inadvertently incentivize unethical behavior, such as manipulating algorithms for personal gain, raising critical questions about the delicate balance between transparency and privacy in algorithmic decision-making. In the medical scenario under consideration, an individual might endeavor to optimize their chances of being chosen for the treatment while simultaneously diminishing the prospects for others.

However, this is not the only problem. Given the advent of novel algorithms in the medical field adept at reestablishing the identities of individuals linked to data, the full disclosure of these algorithms may yield adverse outcomes [18]. Providing tools accessible to all that have the potential to breach individuals' privacy constitutes conduct that is clearly opposed to ethical principles.

This situation highlights the intricate challenge of ensuring that transparency doesn't become a blueprint for misuse, while still maintaining user trust and protecting privacy.

In conclusion, though opinions in privacy-related topics might vary, explainability and transparency are vital to tackle bias issues and develop algorithms that act aligned with ethical behavior.

B. Finances

The use of data science in the loaning sector is not recent. Unlike other businesses, the risk of loaning is elevated and companies try to acquire as much data as possible to elaborate profiles of the demanders, which serve as discriminators[1]. However, these profiles tend to be based on highly discriminatory data. In the 1930s, before the era of artificial intelligence, the lending industry in Chicago engaged in a discriminatory practice known as redlining [19]. This involved categorizing neighborhoods as suitable or unsuitable for mortgage loans based primarily on the racial and economic makeup of their residents. Predominantly Black and lower-income neighborhoods were marked in red on maps, indicating they were high-risk and not worthy of investment. This practice was not based on individual financial situations but rather on broad, prejudiced assumptions about entire communities.

Nowadays, massive amounts of data can be fetched from different sources, which, together with the increase in computational capacity has made profiling more potent, however, it has also made social biases more uncontrollable. The biases in AI banking are starkly illustrated by disparities in FICO® Scores [20], which summarize an individual's credit risk into a three-digit number and are used in 90% of lending decisions. Data reveals a significant racial gap: over 70% of white households in the U.S. have a FICO Score above 700, compared to only 20% of Black households. Additionally, one in three Black households with credit histories has insufficient credit, resulting in no credit score, almost double the rate of white households (17.9%).

Lenders have a fundamental interest in mitigating the risks associated with loans to ensure the survival and profitability of their business. When someone with a high risk of default applies for a loan, the lender must consider mechanisms to reduce this risk. This might include using AI algorithms to better assess the applicant's creditworthiness or requiring collateral. The ultimate goal is to balance credit availability with the financial security of the lender. Nevertheless, while the purpose of AI banking is to move towards a more profitable business model by minimizing risk, it lacks some ethical considerations. Profit is an objective good, however, in light of the inequalities it introduces to society, it might no longer be justifiable.

Contrary to the stereotype of poor financial management, there are many legitimate reasons why a person might need a loan, even if they are considered high-risk

[1]Discriminators in this context refer to tools or algorithms designed to distinguish between different items, categories, or signals, and do not imply social discrimination or unfair treatment.

such as debt consolidation, home improvements, emergency expenses, vehicle financing, moving costs, weddings, or vacations. Going even further, loans play an important social role by providing immediate access to funds that must be repaid later. In an egalitarian society, everyone should have access to a loan, regardless of their initial financial situation. Otherwise, it promotes the stagnation of poverty, as those with fewer resources cannot improve their financial situation without access to credit. [21]

Another issue AI banking presents is the lack of explainability, this one might not get an explanation that dives deep into the reasons why a loan has been denied. These biases in finance are monitored by the CFPB [22], which protects the rights of users to receive detailed explanations of decisions made by artificial intelligence algorithms. Some companies are trying to align their algorithms with more ethical principles such as equality, for example, Zest AI is trying to comply with the CFPB guidelines [23] but acknowledges that "it is not easy" and that "the concern about bias is recent". This company promotes the use of non-generative and non-dynamic artificial intelligence. These models are updated in a very controlled manner, and the predictions are supervised.

In conclusion, although bias in loaning is not a recent event, artificial intelligence has increased its significance and negatively affected the explainability of the decisions of lending companies. Unless there is some legislation, companies could use the potential of AI to mitigate their risks even further, at the expense of equity. However, some companies are trying to align that potential with ethical values, hoping AI will bring ethical fairness some lose blinded by their personal interests.

C. Recommender Systems

In recent years, the level of antagonism between political groups has intensified, with supporters of tione party increasingly disliking and avoiding interactions with members of the opposing party. One of the main culprits identified is social networks. The formation of homophilic social networks, where people preferentially connect with those who share many common connections, has propelled the creation of echo-chambers [24].

Recommender systems filter content to suggest tailored items to users, and several biases arise in them, including popularity bias and unfairness, which are both amplified by the feedback loop these systems rely on.

In the light of the case of Cambridge Analytica [25], research has been conducted on recommender systems during elections, and the European Parliament has adopted new rules that regulate online political advertising during elections in order to ensure the users a secure and reliable digital society. [26] The new regulations seek a singular common regulatory framework to enhance the transparency of sponsored political advertising (both online and offline), reinforce the integrity of election campaigns, and countervail action against disinformation and foreign intervention. The values that are at stake are those of privacy, transparency, and autonomy.

While personalized interactions enhance the user experience on social networks, there is a limit. Users want autonomy and seek to make independent choices without undue influence. From a deontological perspective, these algorithms undermine user autonomy by influencing decisions and perceptions in non-transparent ways. Often, users are unaware of how the content is specifically selected for them, which becomes particularly concerning in the context of political campaigns, as the collected personal data is used to facilitate ad micro-targeting, which can manipulate voting behavior. This practice not only undermines the integrity of the electoral process but also erodes trust in both the platform and society as a whole.

From a philosophical point of view, Karl-Otto Apel highlights how communication can be undermined when one party exploits or manipulates the discourse for its interests without regard for the ethical implications or the broader societal consequences. [27] In recommender systems, addressing these issues requires a heightened awareness of the biases and a commitment to greater transparency and accountability in their design and implementation. Only then, the integrity of our democratic processes and the public trust in social networks will be safeguarded.

Thus, transparency is key in order to ensure a fair environment in social networks and promote a more inclusive and reflective political dialogue, as well as raise awareness of fake content and the existence of echo-chambers.

V. Methodology

Our research aimed to investigate the ethical dilemmas arising from biased AI systems. We began by analyzing key articles and conducting extensive bibliographic research to identify critical issues and establish a foundation for our study. This research provided a comprehensive understanding of biased systems and the previous case analyzes, which we aimed to transfer to our classmates.

To achieve this, we administered an initial questionnaire to assess their prior knowledge in the field, followed by a lesson on the ethical dilemmas related to biased AI. After the lesson, we administered a second questionnaire to observe what they had internalized and the conclusions and interpretations that emerged. Ultimately, we transformed these findings into a list of recommendations and guidelines for becoming a responsible data scientist who consistently considers the potential implications of bias in their algorithms.

A. Questionnaire design and administration

The main objective of the questionnaire is to evaluate the awareness of bias in data and algorithms. [28] It included both direct questions on general knowledge and awareness, together with specific questions for each of the three cases.

- **Medicine:** questions on the limitations and regulations required for data-driven models for medical cases, and the distribution of responsibility and culpability of a decision made through an AI agent.
- **Finance:** Identification of biases that may be introduced to algorithms in finance.
- **Recommender system:** questions on the feed recommended, their interaction with recommended content, and the type of political content shown.

Some questions on regulations and transparent and explainable models allowed for a first evaluation of the preconceived notions and opinions of the surveyed people. Their agreeableness to revealing features, data sources, model architecture, and open-source models was also queried.

Finally, it is noteworthy that nearly half of the individuals responded negatively to the question "Have you ever encountered a biased system?" This suggests a general lack of awareness among a significant portion of people about the algorithms that permeate their lives, how these algorithms function, and the ways in which they can be biased. Consequently, raising awareness and fostering understanding of these issues became a central focus.

B. Conducting a lesson

Following an introduction to a class of approximately 50 students enrolled in the Data Science and Engineering degree program (GCED) at the Polytechnic University of Catalonia (UPC), which encompassed an exploration of data types in AI systems through a variety of illustrative examples and examined the significance of transparency and explainability within the same domain, a subsequent session was conducted. This session aimed to provoke, through a made up case, a more profound thought among the students regarding the intricacies of bias in AI, alongside the crucial elements of transparency and explainability.

The purpose of the session was to equip students with the necessary tools to analyze a case from a philosophical perspective, focusing on the ethical dilemmas discussed in Section 3. With the debate on the case, we expected students to ponder about the different biases that arose from it as well as analyzing the responsible and the guilty and the relation between these concepts and transparency or explainability either in the positive or negative context. The case that was presented can be found in the blog [29].

After presenting the case we opened a round of questions and when all were answered we joined them in groups of 6 and aimed for discussion. During the activity, we used material we had previously prepared to ensure we covered the topics from Section 5.

C. Limitations and challenges

Our study faced limitations due to the homogeneity and size of our sample, which consisted of about 50 third-year Data Science and Engineering students. This lack of diverse backgrounds limited our ability to fully explore the ethical implications of bias in AI systems, as examining such a multifaceted issue requires a wide range of viewpoints. All participants were primarily from the field of system design, which restricted the debate to a single perspective. Engaging with experts from other disciplines would have enriched the discussion, providing a broader range of arguments and insights beyond the participants' usual knowledge frameworks. Additionally, the short time frame and limited prior knowledge on biased AI ethics hindered the depth of our analysis, resulting in less comprehensive conclusions. Despite recognizing these biases, we were unable to mitigate them within the constraints of our study.

VI. Results and Interpretation

The classroom debate yielded arguments and conclusions that demonstrated a profound evolution in the understanding of the discussed topics. All groups engaged in thoughtful ethical reflections on the consequences of integrating biased systems into society, particularly in the healthcare and financial sectors.

Participants identified and examined biases inherent in algorithms, acknowledging their prevalence in daily life. They understood that these biases could arise from data inadequacies, human biases during development, data collection methods, and economic interests. The discussions explored the complexities of biases in algorithmic decision-making and the ethical implications of prioritizing profits and interests over societal needs.

The groups recognized that technology is not neutral. While it may be designed to improve society and simplify life, it can also exacerbate societal inequalities, necessitating reflection on its consequences. Despite recognizing the necessity of demographic or sensitive data in decision-making, they emphasized the importance of filtering data to prevent discrimination.

There was a consensus on the shared responsibility to ensure ethical algorithm use, underscoring the need for transparency and explainability to promote fair and ethical decision-making processes. The limitations to transparency were also reached and discussed, with different points of view emerging. The participants concluded that, as data engineers, it is imperative to minimize harmful biases by using representative data and avoiding harmful stereotypical assumptions.

Moreover, it is worth noting that after participating in the proposed activity, the majority of participants who had initially indicated in the questionnaire that they had never encountered a biased system changed their response. This suggests a significant shift towards a more nuanced and realistic comprehension of the discussed topic, which was one of the main objectives.

VII. Recommendation and guidelines

In the field of data science, ethical practice is crucial. As data scientists, we must balance technical expertise with a commitment to ethical responsibilities. Essential

recommendations and guidelines are explained below to help data scientists navigate ethical challenges, focusing on addressing bias and enhancing transparency and explainability. Key principles include embracing post-conventional ethics, responsibly implementing algorithms, confronting and mitigating bias, and fostering multidisciplinary collaboration.

First, adopting a post-conventional ethical perspective is essential to excel as a data scientist. This means that we must question our actions and their implications, acknowledging that while completely eradicating bias is often impossible, we must stay alert to its presence and try to detect it, ensuring it does not influence our conclusions. Following on this topic, when evaluating the potential impacts of some implementations, it is extremely important to detect the purpose of those. Everything is done with a finality, and we must be sure that it is right and positive. By doing so, we ensure that our work aligns with our society's values and contributes positively to it.

As already stated, a really important aspect is managing and reducing bias. Since we know that it can emerge at various stages of a data scientist project, we must be aware and keep in mind the following guidelines when trying to deal with it:

- **Bias in data:** we should ensure that the datasets are diverse and representative to avoid perpetuating existing inequalities, and we should also be transparent in the data cleaning process, always documenting any exclusions or transformations to avoid introducing new biases. We must also be aware of the inherent biases present in our society. We should work with statistical techniques or processing methods with the objective of reducing bias in our data that has not been previously addressed. These techniques include, for example, data division in subgroups depending on mutual characteristics to determine that all groups are equally represented, also known as stratified sampling.
- **Algorithmic bias:** it refers to the systematic and repeatable errors that create unfair outcomes for certain groups of people. This issue is exacerbated when inherent human biases are reflected in the algorithms, as it is more difficult to identify and mitigate these biases. Using explainable and interpretable models can help reduce it, so whenever feasible, we should opt for algorithms and models meticulously crafted to mitigate bias in our data and promote equity. This is exemplified by the field of Fairness-Aware Machine Learning [30], which, rather than solely prioritizing performance optimization, aims to minimize the model's societal impact by explicitly incorporating fairness criteria during the training and evaluation phases. Additionally, sensitivity analyses are key in this endeavor to reduce bias. By systematically adjusting the model's parameters, we can scrutinize the resultant impacts to identify and address biased variables or techniques embedded within the model.
- **Socio-technical biases:** refers to the inherent prejudices that can be embedded in technology due to social and technical factors, which can manifest during the design and development stages of technology. It is essential to assemble diverse development teams to ensure a wide range of perspectives and experiences. Engaging with a broad spectrum of users through participatory design can help ensure the technology meets the needs of various user groups. Additionally, it is important to be culturally sensitive and consider the local context in which the technology will be used, tailoring solutions to address specific cultural and societal needs. These guidelines can help mitigate socio-technical bias and create technologies that are more inclusive and equitable, ultimately contributing positively to society.

When biases can not be fully controlled, transparency and explainability are critical in order to break the Black Box. Data scientists should be able to articulate the limitations, assumptions, and potential biases of models. Providing comprehensive documentation of the decision-making process and ethical considerations is important too.

Finally, we must know that we are not alone when treating biases, as this work requires collaboration across disciplines. We should foster open dialogues, encouraging discussions about bias and ethics within the organization and with the public. Working with professionals from fields such as ethics, sociology, psychology, and law provides diverse perspectives and a more comprehensive approach to complex ethical issues. This collaborative effort enhances the robustness and fairness of our methodologies and outcomes.

VIII. Conclusions

As we have seen, biases in AI are a significant problem. While we have the necessary skills to understand and address them appropriately, the problem lies in the lack of understanding of biases, which is a generalized problem that negatively affects our society. In our case, this lack prevents us from identifying biases, thus perpetuating and exacerbating the problems they cause.

For that reason, we must raise awareness of biases and try to extend it. As data scientists, we must reflect on the responsibility that comes with our work, embracing a journey of continuous questioning, learning, and ethical improvement. Every action we take, from writing lines of code to training models, can have a significant impact on people's lives. Therefore, it is crucial to consider the ethical and social implications of our decisions and strive to minimize any negative effects that may arise.

Despite efforts, it is important to recognize that biases can persist, perpetuating inequalities and disadvantages. Although it is no longer in our hands, we must not stop here. While complete understanding and explanation of AI systems may be challenging given their complexity and quick improvement, explainability should remain as a

goal to reach. In the meantime, what we can do is prioritize transparency and openness in our work with AI. So despite arguments against creating transparent algorithms due to social or economic benefits, as responsible data scientists we should always value transparency and demand it, especially in delicate cases where it is crucial for the well-being of society.

REFERENCES

[1] Pasquale, F. (2015). The Black Box Society: The Secret Algorithms That Control Money and Information. Harvard University Press.

[2] Lopez, P. (2021, December 8). Bias does not equal bias: a socio-technical typology of bias in data-based algorithmic systems. Internet Policy Review. *https://policyreview.info/articles/analysis/bias-does-not-equal-bias-socio-technical-typology-bias-data-based-algorithmic*

[3] Baeza-Yates, R. (2018). Bias on the web. Communications of the ACM, 61(6), 54–61. *https://doi.org/10.1145/3209581*

[4] Winick, E. (2022, June 17). Amazon ditched AI recruitment software because it was biased against women. MIT Technology Review. *https://www.technologyreview.com/2018/10/10/139858/amazon-ditched-ai-recruitment-software-because-it-was-biased-against-women/*

[5] Marinucci, L., Mazzuca, C., & Gangemi, A. (2022). Exposing implicit biases and stereotypes in human and artificial intelligence: state of the art and challenges with a focus on gender. AI & Society, 38(2), 747–761. *https://doi.org/10.1007/s00146-022-01474-3*

[6] Kohlberg, Lawrence. (1981). The Philosophy of Moral Development. Moral Stages and the Idea of Justice. San Francisco, CA: Harper & Row Pubs.

[7] McGoey, L. (2012). The logic of strategic ignorance. British Journal of Sociology, 63(3), 533–576. *https://doi.org/10.1111/j.1468-4446.2012.01424.x*

[8] responsibility. (n.d.). Oxford Reference. *https://doi.org/10.1093/oi/authority.20110803100416932*

[9] Roscher, R., Bohn, B., Duarte, M. F., & Garcke, J. (2020). Explainable machine learning for scientific insights and discoveries. IEEE Access, 8, 42200–42216. *https://doi.org/10.1109/access.2020.2976199*

[10] Special theme: eXplainable AI. (2023, July). ERCIM NEWS. *https://ercim-news.ercim.eu/images/stories/EN134/EN134-web.pdf*

[11] Saeed, W., & Omlin, C. (2023). Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities. Knowledge-based Systems, 263, 110273. *https://doi.org/10.1016/j.knosys.2023.110273*

[12] Sample, I. (2017, November 27). Computer says no: why making AIs fair, accountable and transparent is crucial. The Guardian. *https://www.theguardian.com/science/2017/nov/05/computer-says-no-why-making-ais-fair-accountable-and-transparent-is-crucial*

[13] Radiology, E. S. O. (2022). Current practical experience with artificial intelligence in clinical radiology: a survey of the European Society of Radiology. Insights Into Imaging, 13(1). *https://doi.org/10.1186/s13244-022-01247-y*

[14] Clinical practice algorithms. (n.d.). MD Anderson Cancer Center. *https://www.mdanderson.org/for-physicians/clinical-tools-resources/clinical-practice-algorithms.html*

[15] Lysaght, T., Lim, H. Y., Xafis, V., & Ngiam, K. Y. (2019). AI-Assisted Decision-making in healthcare. Asian Bioethics Review, 11(3), 299–314. *https://doi.org/10.1007/s41649-019-00096-0*

[16] Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447–453. *https://doi.org/10.1126/science.aax2342*

[17] Wu, C. (2024). Data privacy: From transparency to fairness. Technology in Society, 76, 102457. *https://doi.org/10.1016/j.techsoc.2024.102457*

[18] Murdoch, B. (2021). Privacy and artificial intelligence: challenges for protecting health information in a new era. BMC Medical Ethics, 22(1). *https://doi.org/10.1186/s12910-021-00687-3*

[19] Redlining. (n.d.). *http://www.encyclopedia.chicagohistory.org/pages/1050.html*

[20] Explaining the Black-White homeownership gap: A Closer Look at Disparities across Local Markets. (2019). URBAN INSTITUTE. *https://www.urban.org/sites/default/files/publication/101160/explaining_the_black-white_homeownership_gap_a_closer_look_at_disparities_across_local_markets_0.pdf*

[21] Valenzuela, P., & Bonilla, Á. (2015). La pobreza y el crédito: Entre la inclusión y la vulnerabilidad. Dialnet. *https://dialnet.unirioja.es/servlet/articulo?codigo=6310252*

[22] Forensic & Integrity Services, Ernst & Young LLP, How to mitigate AI discrimination and bias in financial services. *https://www.ey.com/en_us/insights/forensic − integrity − services/ai−discrimination−and−bias−in−financial−services*

[23] Zerucha, T. (2023, October 5). AI, if used properly, can improve loan decisioning. Fintech Nexus. *https://www.fintechnexus.com/ai-if-used-properly-can-improve-loan-decisioning/*

[24] Spohr, D. (2017). Fake news and ideological polarization. Business Information Review, 34(3), 150–160. *https://doi.org/10.1177/0266382117722446*

[25] History of the Cambridge Analytica controversy. (2023, March 16). Bipartisan Policy Center. *https://bipartisanpolicy.org/blog/cambridge-analytica/*

[26] EUROPEAN UNION. (2024, February 29). REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on the transparency and targeting of political advertising. *https://data.consilium.europa.eu/doc/document/PE-90-2023-INIT/en/pdf*

[27] Kettner, M. (2011). Discourse Ethics beyond Apel and Habermas. A Realistic Relaunch. Nordicum-Mediterraneum, 6(1). *https://doi.org/10.33112/nm.6.1.4*

[28] TAED1- Bias and Transparency. *https://forms.gle/sVuNrcnZJubzHN7v9*

[29] G, T. (2024, April 15). Second lesson — TAED G2. *https://bias.bla.cat/posts/second-lesson/*

[30] Fairness-aware Learning through Regularization Approach. (2011, December 1). IEEE Conference Publication — IEEE Xplore. *https://ieeexplore.ieee.org/abstract/document/6137441*