CEST
Centro de Estudos Sociedade e Tecnologia
USP
Universidade de São Paulo
Bulletin - Volume 7, Number 07, September/2022

Ethics in Data Science: Reflecting realities and not interests

Giovanna Cavali, João Antônio Silva e Larissa Pereira de Andrade

The definition of ethics refers to the philosophy responsible for investigating the principles that motivate, deform, discipline, or guide human behavior, reflecting on the essence of norms, values, prescriptions, and exhortations present in any social reality. By extension, it is the set of rules and precepts of an evaluative and moral order of an individual, a social group or a society.

Data ethics evaluates data practices, which include the collection, generation, analysis, and dissemination of data, structured and unstructured, and that have the potential to positively or negatively impact individuals and society. Data ethics describes a code of behavior that must be used at all stages of a data science project. About data ethics, there are other definitions:

Data ethics describes the code of behavior that must be used at all stages of data science projects.

- "Data mindfulness" guides the ethical use of data in organizations.
- Not just moral guidance on "what data to collected and how to use them", but also on who makes those decisions in the first place.
- A code of conduct or ethics for Data Scientists, like the purpose of the Hippocratic Oath in guiding medical professionals.
- Code that guides the behavior of Data Scientists towards a "better human society".

Data Science tools, especially machine learning, are increasingly used to support decision-making. Consequently, we see a gradual reduction of human intervention in areas that affect aspects of our daily lives. However, any failure in algorithmic judgment can have significant implications. Therefore, it is essential to define proper guidelines to build reliable and responsible machine learning solutions, taking ethics as a central pillar. Ethical Artificial Intelligence (AI) is a broad subject covering many topics, such as privacy, data governance, social and environmental well-being, algorithmic accountability, and traceability.

As machine learning algorithms and their abstractions and hypotheses become more complex, it is increasingly difficult to understand all the possible consequences of these models. Several examples of unfair and discriminatory machine learning algorithms have made their way into the media in recent years. Some of these examples are:

- COMPAS was a widely used commercial software that measured a person's risk of committing an upcoming crime. The algorithm used was compared to human judgment in one study. It was later found that COMPAS was more likely to assign a higher risk to African American offenders than Caucasians of the same profile.
- Gender bias was detected in early versions of Google Translate.
- Goldman Sachs was investigated for using an AI algorithm that allegedly discriminated against women by giving men higher credit limits on their Apple cards.
- A healthcare algorithm used in hospitals in the United States has discriminated against black patients. The AI was intended to measure which patients would

benefit most from accessing a high-risk health care management program. However, when they compared the AI-generated risk score with other health scores on their patients, they found that black patients were consistently underestimated.

These are just a few examples proving the importance of having a defined process to analyze, identify and mitigate potential problems in designing, implementing and validating AI models. If, on the one hand, the result of the analysis of an AI may reflect problems intrinsic to the data that used in its construction; on the other hand, the result of the analysis can be planned to achieve some objective or mask some reality. In early 2020, the data science community was rocked by the contest cheating scandal by Kaggle (one of the leading communities of data scientists). The competition consisted of developing an algorithm to predict the rate of adoption of animals based on listings on PetFinder.my, a Malaysian website. The "winning" team obtained the test data likely by copying data from Kaggle or PetFinder.my itself, encoded and decoded it in their algorithm to obfuscate their ill-gotten advantage. The winning team was subsequently disqualified.

Another example of malicious use of AI models is deep fakes, videos or images that show a person doing or saying something they didn't do. Recently, a deep fake video of Ukrainian President Volodymyr Zelensky made headlines in several media outlets. In the fake video, Ukraine's president advised civilians to surrender to the Russian military and return to their families, contradicting his previous speeches. Zelensky himself released a statement on Instagram denying the deep fake.

Some examples of problems and complications arising from AI models were mentioned. Still, it is essential to note that new applications are emerging daily, with risks that we do not yet know. For this reason, it is vital to raise this agenda in different areas of society. For example, without caution in implementing and monitoring data science results, all the processes may fail to optimize and reflect reality, reflecting interests in a discriminatory and unfair way

Companies that apply the ethical principles of justice, privacy, transparency and responsibility to their Artificial Intelligence models and in analyzing their results can use them as competitive advantages. In addition to mitigating compliance risks in the use of data, adherence to data ethics practices brings greater trust and loyalty to customers and investors, with positive impacts on the company's reputation and value.

Giovanna Cavali *has a bachelor's degree in Mechanical Engineering at Escola Politécnica da Universidade de Sao Paulo and is MBA-USP student at Data Science and Analytics of the Programa de Educação Continuada da Escola Politécnica da USP*

João Antônio Silva *has a master's degree in Computer Science at Universidade Federal de Lavras and is MBA-USP student at Data Science and Analytics do Programa de Educação Continuada da Escola Politécnica da USP*

Larissa Pereira de Andrade *has a degree in Science -Mathematics at Universidade Federal de São Paulo and is MBA-USP student at Data Science and Analytics do Programa de Educação Continuada da Escola Politécnica da USP.*

Academic Coordinator: Edison Spina

This article is a result of the authors' ascertainment and analysis, without compulsory reflecting CEST opinion.