Metadata

- Id: EU.AI4T.O1.M3.3.3t
- Title: 3.3.3 Where Does the Risk Come From?
- Type: text
- Description: Identify the different types of risks
- Subject: Artificial Intelligence for and by Teachers
- Authors:
  - AI4T
- Licence: CC BY 4.0
- Date: 2022-11-15

WHERE DOES THE RISK COME FROM?

In its study about Artificial Intelligence[1], the European Parliamentary Research Service stated: "*It is important to note that AI algorithms cannot be objective because, just like people, in the course of their training they develop a way of making sense of what they have seen before, and use this 'worldview' to categorise new situations with which they are presented.*"

Let's see where the subjectivity of an AI comes from and what are the associated risks.

THE BIAS IN DATA AND IN ALGORITHMS

As for any digital system, the data used in AI-based platforms come from different sources and have multiple formats. They carry different types of bias[2]. Data bias is mainly statistical. Let's list a few of them.

- **Sample bias** is typically present in data values. For example, this is the case for a recruitment algorithm trained on a database in which men are overrepresented will exclude women.
- **Stereotype bias** is a tendency to act in reference to the social group we belong to. For example, one study shows that women tend to click on job offers that they think are easier to get as a woman.
- **Omitted variable bias** (modelling or coding bias) is a bias due to the difficulty of representing or coding a factor in the data. For example, because it is difficult to find factual criteria to measure emotional intelligence, this dimension is absent from recruitment algorithms.
- **The selection bias** is in turn due to the characteristics of the sample selected to draw conclusions. For example, a bank will use internal data to derive a credit score, focusing on

those who have or have not obtained a loan, but ignoring those who have never needed to borrow, etc.

The algorithmic bias is mainly a matter of reasoning. Such bias is introduced by AI engineers deliberately or not.

The previously mentioned European Parliamentary Research Service study gives two concrete examples: "*Consider a symbolic AI algorithm for examining job applications. It might evaluate candidates by assigning scores only on the basis of their education and experience. Yet, if it fails to take account of factors such as maternity leave or to appropriately recognise education in foreign institutions in ways that human selection committees would, the algorithm might discriminate against women and foreign candidates.*"

"*Now, consider a similar AI tool within the ML (Machine Learning) paradigm. Such algorithms find their own ways of identifying which kind of candidates were selected in their training data. Where there is a history of structural biases in these selections -- for example racial discrimination -- the algorithm can learn these. Even where data about nationality or ethnicity is removed from the data, ML is adept at finding proxies for underlying patterns in other data such as languages, postcodes or schools that can be good predictors of ethnicity.*"

THE THREE FACETS OF ALGORITHMIC RISK

The algorithmic risk can be characterised in three ways[3].

- Firstly, there is **algorithmic confinement**, which can also relate to opinions, cultural knowledge or even commercial practices. Indeed, the algorithms confront the Internet user with the same content, depending on his profile and the integrated parameters, despite the respect of the principle of fairness. This is the case on news recommendation sites such as Facebook or product recommendation sites such as Amazon.
- The second facet of algorithmic risk is linked to the **control of all aspects of an individual's life**, from the regulation of information for investors to his or her eating habits, hobbies, or even health status. This tracing of the individual suggests a form of surveillance that contravenes the very essence of individual freedom.
- The third is related to the **potential violation of fundamental rights**. In particular, algorithmic discrimination defined as unfavourable or unequal treatment, in comparison with other persons or other equal or similar situations, based on a ground expressly prohibited by law. This encompasses the study of the fairness (*fairness*) of ranking (sorting of people looking for a job online), recommendation, and prediction learning algorithms. The problem of discriminatory bias induced by algorithms concerns several areas such as online hiring, court decisions, police patrol decisions, or school admissions.

HOW TO DEAL WITH DATA AND ALGORITHMIC RISKS?

For R. Schwartz & al.[4], "*Bias is neither new nor unique to AI and it is not possible to achieve zero risk of bias in an AI system*".
Meanwhile, recognizing that AI agents are inherently subjective is a crucial prerequisite for ensuring that they are only applied to tasks for which they are well equipped.

EPRS' study concludes with several recommendations when using AI-based applications:

- Understand bias and subjectivity
- Avoid applications beyond AI's capabilities
- Avoid applications with undesirable impacts
- Maintain human autonomy
- Look for solutions to problems, not problems for solutions
- Consider what we really want from AI

---

[1] Artificial intelligence: How does it work, why does it matter, and what can we do about it ? - Philip Boucher, Scientific Foresight Unit (STOA) - ISBN: 978-92-846-6770-3 - Union Européenne, 2020 ↩

[2] Algorithms, Data and Bias: Public Policy Needed, Anne Bouverot, Thierry Delaporte, 2019 ↩

[3] Article in French: D'où vient le risque ? Des données et des algorithmes - Serge Abiteboul, Thierry Viéville, 2020 ↩

[4] Towards a Standard for Identifying and Managing Bias in Artificial Intelligence - Reva Schwartz, Apostol Vassilev, Kristen Greene, Lori Perine, Andrew Burt, NIST Special Publication 1270 , 2022 ↩