GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

# The ethical challenge of working with biases

## From cognitive biases to intersectionality and why it matters for software development

Biases can be defined as miscalculations, distortions in the distribution of training data of artificial-intelligence-algorithms or cognitive limitations like stereotypes. Biases are an ethical challenge because they support the discrimination and unfair treatment of certain individuals or groups based on their race, gender, age, religion, sexual orientation, or other personal characteristics.

Did you ever wonder why some ads for high paying jobs get displayed to men and not women ?[1] Why some people are paying more for insurance than others ?[2] How it happens that images created by AI show women sexified and men as astronauts ? [3]

Those are examples for the representation of biases in algorithmic systems. When biases influence decision-making in areas such as hiring, promotion, or access to resources, they create systemic barriers that limit opportunities and perpetuate social inequality. In addition programmers might encode their biases into their programming artefacts (Cowgill et al 2020) and reinforce discriminatory attitudes and behaviours with the resulting algorithmic systems. To address biases, it is important to raise awareness about their existence and impact, to examine the underlying causes and contexts in which they occur, and to develop strategies and tools to counteract them.

> “Algorithm development implicitly encodes developer assumptions that they may not be aware of, including ethical and political values. Thus it is not always possible for individual technology workers to identify or assess their own biases or faulty assumptions.” (Raji et al 2020)

## Social biases

Social biases refer to systematic and often unconscious ways in which individuals hold and act on preconceived beliefs, attitudes, and stereotypes about others based on social categories, such as race, ethnicity, gender, age, sexual orientation, religion, and socioeconomic status, and more. Social biases may manifest in both implicit and explicit forms. Implicit biases are unconscious associations between social categories and certain qualities or behaviours, while explicit biases are conscious and intentional beliefs or attitudes that individuals hold about others. These biases are influenced by personal experiences, cultural and societal norms, and social conditioning.

Any Classification of people into groups like “wealthy”, “fit” or “trustworthy” is a social construct. While stereotyping helps to categorize new information quickly, it should not be treated as valid information. Any description of what counts as normal is an assumption that would need to be verified with actual people. (Gasser et al. 2020)

[1] Datta, A., Tschantz, M. C., and Datta, A. (2015). Automated Experiments on Ad Privacy Settings. Proceedings on Privacy Enhancing Technologies, 2015(1)

[2] https://www.theguardian.com/money/2018/oct/31/insurance-regulator-look-possible-racial-bias-financial-conduct-authority

[3] https://www.technologyreview.com/2022/12/12/1064751/the-viral-ai-avatar-app-lensa-undressed-me-without-my-consent/

## Cognitive biases

Cognitive biases appear in the human brain when it is analysing information. This means when processing information that is too voluminous or too complex for the human brain to handle, it will find shortcuts. So when you are forced to make a rapid judgment in a time-frame that is too short to review the information at hand, your brain will still provide you with a possible answer. When there is insufficient information for making the decision, you are still able to guess. The following three examples show different types of biases, to get a further insight into the many types of cognitive biases see the cognitive bias codex graphic.

**Confirmation bias:** This is the tendency to seek out and interpret information in a way that confirms your existing beliefs or expectations. This can lead to the confirmation of incorrect assumptions or the failure to consider alternative solutions or approaches. This is our tendency to search for and recall information that confirms our existing beliefs or hypotheses. A classic example of this in performance analysis is, “this component has never been a problem in production, so we know it is not the root cause.” The correct conclusion would be, “we expect this component not to be root cause.”

**Anchoring bias:** This is the tendency to rely too heavily on the first piece of information encountered when making decisions. This can lead to overreliance on early assumptions or requirements, which can prevent from considering changing circumstances or overcoming the influence of irrelevant or misleading information. The anchoring effect is the tendency to rely too heavily on an initial piece of information, known as the anchor, when making decisions. This is used in shopping, when the intial price is used as an anchor against which the sales price will be judged.

**Sunk cost fallacy:** This is the tendency to continue investing in a project or decision, despite evidence that it is not working, because of the amount of time or resources already invested. This can lead to a reluctance to change direction or abandon unsuccessful projects, and a failure to explore how to make more rational and adaptive choices.

## Technical biases

Technical biases can emerge by using technology in a specific way. For example when a program assumes causal relations between related variables or databases can only provide the information that is indexed. (Noble 2018) Other bias show when the source data for algorithmic systems is biased due to the absence of specific information or is reflecting historical biases. Sample-data that the algorithms are trained on are usually not sufficiently representative of the larger population-data that the algorithms are subsequently used in. Technical constraints arise from the software and the hardware that is used for the computers running a program. On the hardware side sensors, batteries, or possible motor movements restrict what a machine can do. On the software side these could be certain defined input formats or information loss while transferring information between different programming languages. They also addressed as a side-effect of deliberate decisions like choosing a ranking algorithm, a specific set of sensors or cost saving measures. Last but not least depending on context of deployment a system might be viewed as fair in some circumstances but not in others.

## Intersectionality

Intersectionality is a concept that was introduced to address that concepts of discrimination failed to account for the unique experiences of individuals who belong to multiple marginalized social groups. (Crenshaw 1991) Crenshaw emphasized that experiences of discrimination cannot be fully understood by examining just one aspect of an individual's identity. For instance, a woman of color may face a different form of bias compared to a white woman or a man of color. Intersectionality provides a conceptual lens to understand discrimination not only on one axis but also on the intersection of multiple social categories. The concept of intersectional bias is necessary because it recognizes that individuals' experiences of bias and discrimination are influenced by multiple factors that intersect and interact. Automatic tests for specific biases have to account for the fact, that some discrimination will be more prevalent on the intersectional level. (Sweeney 2013)

## To what extent can we avoid to let biases influence our own decisions ?

A common effect associated with the presence of biases is the so-called biases blindness. It describes the tendency that most people consider themselves uninfluenced by biases. Above biases show that everybody is influenced by their own knowledge and experiences. Awareness will reduce the effects but reducing bias is a group effort. Diverse teams can create environments with different biases. So teams can discuss and reach new insights and information before they reproduce biases in software structures.

Being human means thinking with a brain prone to skipping contradicting information and sometimes you are working without time to research other options. While there is no way to completely lose the possibility for biases communication in diverse teams and learning from feedback are possible options to reduce harmful outcomes.

## Activities

Play a game that presents you with your own biases https://www.happybrainscience.com/blog/how-to-use-a-fun-party-game-to-tackle-implicit-bias/

To see how many likely falsehoods have already been identified check this selection on github https://github.com/kdeldycke/awesome-falsehood

## How do we recognize our own biases in our project?

## How can we create structures that counteract biases in our project?

## Further reading

- Thea Gasser, Eduard Klein, Lasse Seppänen, Bias – A Lurking Danger that Can Convert Algorithmic Systems into Discriminatory Entities, (2020) https://arbor.bfh.ch/13189/1/Bias_Gasser-Klein-Sepp%C3%A4nen_centric_2020_1_10_30004.pdf

- Dr. Safiya Umoja Noble, Algorithms of Oppression,

- Fabrizio Dell'Acqua, Bo Cowgill, Samuel Deng, Daniel Hsu, Nakul Verma, Augustin Chaintreau. Biased Programmers? Or Biased Data? A Field Experiment in Operationalizing AI Ethics (2020) Columbia Business School and Columbia University Department of Computer Science

- Kimberlé Crenshaw - "Mapping the Margins: Intersectionality, Identity Politics, and Violence against Women of Color" (1991) https://doi.org/10.2307/1229039

TED Talk: https://www.ted.com/talks/kimberle_crenshaw_the_urgency_of_intersectionality

- Latanya Sweeney - Discrimination in Online Ad Delivery (2013) https://dx.doi.org/10.2139/ssrn.2208240

- Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, Parker Barnes - Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing (2020). https://doi.org/10.48550/arXiv.2001.00973

- Drew Roselli, Jeanna Matthews, and Nisha Talagala. 2019. Managing Bias in AI. In Companion Proceedings of The 2019 World Wide Web Conference (WWW '19). Association for Computing Machinery, New York, NY, USA, 539–544. https://doi.org/10.1145/3308560.3317590

- Johansen, J., Pedersen, T. & Johansen, C. Studying human-to-computer bias transference. *AI & Soc* (2021). https://doi.org/10.1007/s00146-021-01328-4

- Cognitive Bias Cheat Sheet https://github.com/busterbenson/public/blob/master/cognitive-bias-cheat-sheet.json

- Patricia Hill Collins - "Intersectionality's Definitional Dilemmas" (2015) https://www.annualreviews.org/doi/abs/10.1146/annurev-soc-073014-112142

- Angela Davis - "Women, Race, and Class" (1981)