Artificial Intelligence and Social Computing, Vol. 72, 2023, 181–188
https://doi.org/10.54941/ahfe1003286

Lowering the Risk of Bias in AI Applications

Jj Link, Anne-Elisabeth Krüger, and Helena Dadakou

Fraunhofer IAO, Fraunhofer Institute for Industrial Engineering IAO, Stuttgart, Germany

ABSTRACT

Data is not free of biases, and AI systems that are based on the data are not either. What can be done to try the best, to minimize the risk of building systems that perpetuate the biases that exist in society and in data? In our paper we explore the possibilities along the Human Centered Design Process and in Design Thinking, to lower the risk of keeping imbalances or gaps in data and models – marking and discussing relevant checkpoints along the process, which are prone to bias. But looking at the design process is not enough: Decision makers, development team and design team, respectively their composition and awareness towards risks of discrimination and their decisions in involving potential users and non-users, collecting data and testing the application also play a major role in trying to implement systems with the least biases possible.

Keywords: AI, Artificial intelligence, Anti-discrimination, Biases, Critical diversity, Diversity, Human-centered design process, Design thinking, Ml, Machine learning

INTRODUCTION

Data Assets and Technology are not neutral. They reflect the situated knowledge we inscribe into them, our perspectives and biases. Especially AI Applications that rely heavily on data. But what can we do, to lower the risk of biases in AI Applications?

Having started with a workshop held in November 2021 at World Usability Day and repeated about a year later, we explore the design decisions being made along the development process of an application, and try to find checkpoints, for what would be good premises for these decisions.

Knowledge and Bias

How is it possible that biases, prejudice and inequalities are being reinforced by or reiterated in technology? The main aspects that may contribute to potentially discriminating systems are data containing biases, prejudice or gaps of ignorance, and the human factor of the persons involved in design and development. As a third area, which contains potential for positive or negative contribution to biases are the methods being used in the design process.

Data and technology are forms in which human beings store and transport knowledge. The myth of objectivity and neutrality of math and data still exists in science and economy, although since about 35 years the concept of ‘situated knowledge’ has been around (Haraway, 1988). Emphasizing, that there is a link of any knowledge to its respective context and to the perspectives on the world of the persons building it (the knowledge), it is clear, that all the structured and non-structured data the current AI (Artificial Intelligence) or ML (Machine Learning) Systems access, is also not free of context. And being man (and sometimes woman) made - thus not free from biases.

In discussing the basics of diversity and discrimination we stress the importance of an anti-discriminatory attitude. Discrimination can be directed towards a variety of factors or diversity features. These can be categorized into inner, outer and organizational dimensions of diversity (Gardenswatz and Rowe, 1995). While direct and open discrimination is visible, indirect, structural and institutional discrimination is often overseen. An intersectional perspective as mentioned by (Crenshaw, 1989) helps to understand how multiple vulnerabilities interact or more than add up.

In this paper we are mainly looking at the Human Centered Design Process, which in theory already contains some good possibilities to make sensible decisions with an awareness for diversity and in an anti-discriminatory attitude: Involving the (potential) users is an effective way of making sure, the system meets their requirements and needs. But there are compromises to be made: Time and money are scarce. So, there are limits to communication with users, to testing and having iterations in the process. And of course, the knowledge and experience of the persons acting matters.

While the Human-Centered Design Process itself is a standardized model aiming to include the users‘ (psychological) needs in the best way possible, it can be argued, that some of the methods used along the way may be prone to containing biases and stereotypes, as for the Persona Method (Marsden, 2014), the risk of systems inheriting biases in data and subjective ratings (Stern 2022) or from prior systems design.

For the Design Thinking Process which is also closely linked to tools and methods of design (e.g., Empathy Maps), there are several papers and articles discussing the potential effects of biases. Pervall (2022) and Banerjee (2018) both discuss several unconscious bias effects, that may influence the design process and how to counter act, mentioning for example the method of Assumption Mapping for Perceptual Bias and planning on enough Resources as a counter measure to Omission Bias.

Lee (2022) suggests using the method of Liberatory Design (Anaissie, 2021) as a variant of Design Thinking, to escape some of the threats. Here, the Power Dynamics between designers and users are mentioned, which we argue go along with a sense of responsibility on the importance of the work done by the design team. Furthermore, Liberatory Design asks for the practise self-reflection and self-awareness: This can be done by studying questions preventing discrimination to be asked during the process for better succeeding in a more inclusive and equitable design.

EXPLORING THE RISKS OF BIAS IN THE DESIGN PROCESS

Based on findings of desk research, experience and observation (two online workshops with usability professionals) in the field of human-centred design,

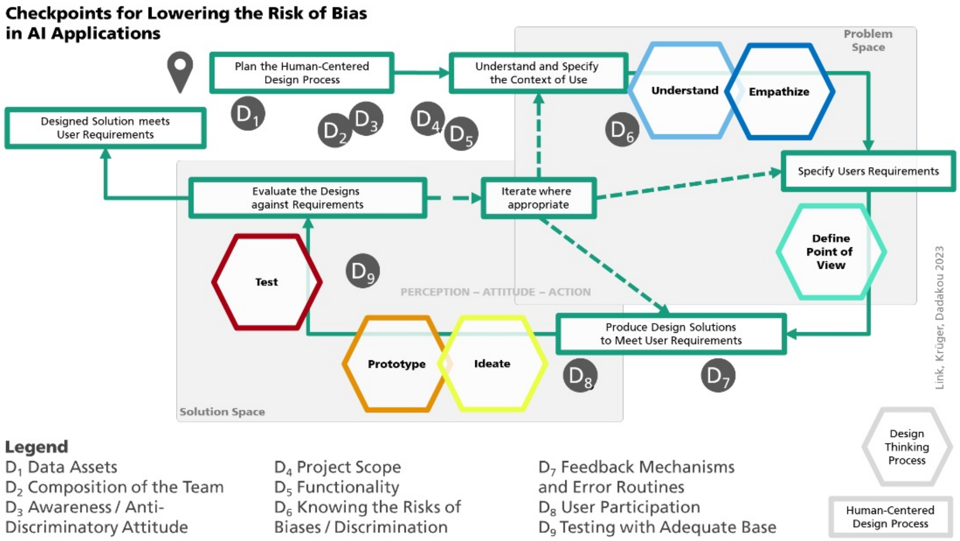
Checkpoints for Lowering the Risk of Bias in AI Applications
Plan the Human-Centered Design Process
Understand and Specify the Context of Use
Understand
Empathize
Problem Space
Designed Solution meets User Requirements
D1
D2
D3
D4
D5
D6
Specify Users Requirements
Evaluate the Designs against Requirements
Iterate where appropriate
Define Point of View
Test
D9
PERCEPTION – ATTITUDE – ACTION
Produce Design Solutions to Meet User Requirements
Prototype
Ideate
D8
D7
Solution Space
Link, Krüger, Dadakou 2023
Legend
D1 Data Assets
D2 Composition of the Team
D3 Awareness / Anti-Discriminatory Attitude
D4 Project Scope
D5 Functionality
D6 Knowing the Risks of Biases / Discrimination
D7 Feedback Mechanisms and Error Routines
D8 User Participation
D9 Testing with Adequate Base
Design Thinking Process
Human-Centered Design Process

Figure 1: At many points in the development process there are design decisions being made. D1-D9 mark the checkpoints, which may heavily influence the potential of bias in an AI application.

we identified the following vital points D1-D9 along the Human-Centred Design Process (ISO 2019) (see Figure 1, rectangular boxes), respectively Design Thinking Process (see Figure 1, hexagonal boxes), at which important decisions regarding discrimination and bias towards AI applications can be made. As recently Design Thinking has become more and more popular in a corporate context, we have mapped the steps of the Human-Centered Design Process to the steps of the Design Thinking Process according to Burghardt (2011).

Each of the checkpoints we have found in our work, in workshops and the following research is marked by a point labelled with D1-9. This list may not be exhaustive, but it contains the relevant points we found until now.

For an easy overview the points are listed here and briefly described. The individual checkpoints are described in detail afterwards.

1. Data Assets
2. Composition of the Team
3. Awareness / Anti-Discriminatory Attitude
4. Project Scope
5. Functionality
6. Knowing Vulnerabilities / Risks of Discrimination
7. Feedback Mechanisms / Error Routines
8. User Participation
9. Testing with an Adequate Base.

The first big limitations to the scope of a project are being made quite at the beginning: What will it be, for whom, what is the goal? (D4). What data is being used? (D1). Who decides and who implements? (D2). At this point it gets obvious, that the knowledge from education and from experience in the team is crucial for the scope of aspects that will be considered (D3). These first definitions early in the project influence also the first definition of functionalities, on which the further requirements analysis might be based (D5). Later in the process, when user requirements are specified – the knowledge on vulnerable groups and various risks of discrimination is important. (D6). A sensitivity towards those aspects can be achieved by having a diverse composition of team members, accompanied by specific education regarding those topics. Thinking of feedback mechanisms and error routines will ensure that users can report on faulty behaviour of systems or make the systems fail gracefully(D7). Involving the right number of users is another key to good coverage of the target group in the ideation process (D8). And finally testing with a basis that is broad enough will make sure, traps are avoided (D9).

Data Assets (D1)

The importance of the structure and quality of data for learning systems has been widely recognized during the last few years (D1), and the words of 'crap in, crap out' is a frequently cited phrase in many projects. But the full number of imbalances that may exist in data is still being explored. Countless examples of poorly functioning systems have been documented and made available to a general public and to a professional audience. One of the most famous examples on social media being a discriminating soap dispense (Afigbo, 2017), many books appeared (Criado-Perez, 2019), (Orwat, 2019), and the studies of Joy Buolamwini make up a big part of the documentary 'Coded Bias' (Boulamwini, 2021). The biases refer to several different characteristics, race and gender being amongst the more obvious examples. The not so obvious parts of less visible and less direct discrimination risks are not less evil. So, is it 'bias in, bias out'?

The good news is, that if we are aware of the lacks in our data, we can try to fix this, and not only have the coded bias in our applications, but also code the countermeasures. For example, Wang (2022) line out in their research that it is not only important which characteristics and data are covered in a system, but also what is the granularity that is used. Schiebinger (2021) suggests fairness evaluations of models in Machine Learning and gives advice on the sub-categories of machine learning. There is still research needed on this topic, but what is maybe more important is spreading the knowledge that is already there.

Composition of the Team (D2), Anti-Discriminatory Attitude (D3), Knowing Risks of Bias and Discrimination (D6)

A general lack of awareness and knowledge on biases and the mechanisms of discrimination that is present in society in general may also show up when we look at the persons working on a project (D2). Of course, there are exceptions, but statistically the persons involved in an AI- or ML-project in Germany, where our observations were made, will be predominantly male, predominantly white, and most of them with an academic grade. Studies show that a similar distribution is likely to be found all over the word (West 2019). Often these factors also bring with them, that these persons involved have a good and regular income, have a good health and have few barriers to reading and writing, are accustomed to a life with much technology and have good access to information.

In short, the persons in such a project team often are privileged in various ways, and not all of them are aware of this. Although it is often a mixed team consisting of a project lead, developers, usability or user-experience professionals, graphic and interaction designers and the purchasers, the group is still quite homogenic in the ways described above.

Generally, in a development team, but in homogenic teams even more, there's danger of implementing via the 'I-Methodology' (Oudshoorn, 2004), meaning that people assume that the later users will think and act like themselves, and that developing a product that works for them will also work for everybody else. Awareness of the existence of this tendency and trying to follow methods to involve users will help not to fall for this.

Additionally, for overcoming the homogeneity in the team, it is a good idea to try to compose the project team in a manner that is a diverse as possible. This reaches out until the hiring policy of the companies involved and the company culture. Composing a heterogenic team that unites people with different perspectives gives the chance that from a variety of possible solutions or combining them a better solution may come up than in a homogenic team.

Another component that is beneficial for the team is general the awareness (D3) the team has for diversity questions and the risk of bias. Basis for a critical attitude towards discrimination is also to know about the risks and mechanisms of discrimination to be able to perceive them, to recognize them and to act on them (D6), the trifold of perception, attitude and action is laid out in (Kinder, 2020). Acting requires background knowledge on the context of the users and empathy. An Empathy Map – a canvas to design a persona, like mentioned earlier, may be a good tool if valid assumptions are being made, while it may stay superficial when people act without a good foundation. Some misconceptions in this phase may be corrected by a well conducted testing procedure (D9), but it is better and probably cheaper to prevent this from the start.

Project Scope (D4) and Functionality (D5)

How important the composition of the project team is, is also obvious when thinking about how the decision for the project is being made: Who decides, that it is worth to invest time and money in the problem that should be solved? Whom will the project be developed for? For whom not? And what is the purpose of the project in general?

Asking these questions at early stages is beneficial for making sure that the project meets its purpose. To help remember those question which in German are called 'W-Fragen', corresponding to 'Wh-questions' in English we will mention here that they are part of the famous title song of the children's show of Sesame Street, underlining that is an elementary thing to stay open and curious to the world around you and ask the questions again and again. A very important question is also 'Who not?' – 'Who ist not here?' or 'Who are not the persons we design this for and why?' More on the 'Wh-questions', and how exclusions and inclusions in the target group and the general scope is described in (Kinder, 2020). This is relevant when defining the functions of the solution to be developed on a macro and micro level. User Research by a team which is aware of the pitfalls around deciding to fast, will help to build a requirements analysis and a comprehensive idea on the groups of users and the non-users on a stable basis.

Feedback Mechanisms (D7) and Error Routines (D7)

A very good question while developing a system, especially a learning system, is to ask: 'How can I tell the system that I was not satisfied with the outcome of my interaction?' We asked this question in our workshop, and amongst the solutions the participants came up with there were suggestions that stayed within the context of use, like for example providing feedback on a feature within the same interface, and suggestions that used different ways like calling the customer service or posting a video of the non-working solution on social media. If it can be foreseen that a feature may have difficulties it is wise to provide a channel for feedback.

And while ideally errors should be prevented or be easily undoable (Nielsen 1994), there is a certain chance that even if you plan for the worst 'What could possibly go wrong?'. User participation and testing with a lot of attention for any kind of irritation will help (D7). In many cases a lack of time and money and maybe pessimism will prevent the development team from fully diving into the possibilities here. But the chance of developing learning systems is a good opportunity to think about new possibilities at this point.

User Participation (D8) and Adequate Testing Base (D9)

Involving a good number and a representative part of users into the (co-)creation process will help to come up with good design solutions (D8). Pitfalls here may lie in the selection or self-selection of users. This also applies for the tests (D9). Here again there's a good 'W-question' to ask: 'Who is not here? Why?' Will the way we set up the co-creation or testing phase exclude a group that is relevant? By not inviting them, by maybe only providing time slots, that persons who take care of children are unlikely to be available, by not presenting our ad to people outside our own bubble or a big part of our target group being just too busy by earning their living to take part. The place where we conduct this creative or testing process will produce exclusions and inclusion and the language we write in. From our experience we can say: Going to where your users are is always a good idea.

CONCLUSION

The project definition (D4) and finding funding for an AI or ML project relying on data (D1) that may or may not be biased are the first steps, where relevant decisions are being made. Getting a team (D2) that brings together various perspectives (D6) early in the design process is helpful with choosing from or prioritizing several ideas, as well as it is important for setting up the plan for implementation and functionality (D5). Ideas like gender budgeting or other quota may help to make sure the perspectives of marginalized groups are considered in an appropriate way. It is obvious that persons with a heightened awareness (D3) will be able to understand the context of use more easily find the right vocabulary to describe their findings. People who do not share the experience of being discriminated against may have a harder time in recognizing vulnerabilities and keeping them in mind in latter phases, like when thinking about what could go wrong (D7). Inviting the right persons to share their visions and needs (D8) and to bring usability and user experience issues to the surface while testing (D9) are vital points when developing systems that seem to operate 'only on objective data'.

We hope to provide an accessible and graspable way for usability and user experience professionals and any other person involved in developing data-based systems. Although the depiction of the Human-Centered Design Process and the Design Thinking Process and our checkpoints in one image (Figure 1) is quite a lot in one place – and that's still without having marked all the methods that may be used. But it may be a good canvas to put up next to your desk. To remind you and your fellow members of the project team of learning (perceiving and developing an attitude) and acting on the complexity of interactions and the diversity in human beings in a compact way.

ACKNOWLEDGMENT

The authors would like to thank Peggy Piesche, Katja Kinder, Maisha Auma and Doris Janssen.

REFERENCES

Afigbo, C. (2017). https://twitter.com/nke_ise/status/897756900753891328

Anaissie, T., Cary, V., Clifford, D., Malarkey, T. & Wise, S. (2021). Liberatory Design. http://www.liberatorydesign.com

Banerjee, S. K. (2018). Cognitive Bias in Design thinking https://www.linkedin.com/pulse/cognitive-bias-design-thinking-saranya-kumar-banerjee/

Buolamwini, J. & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In Proceedings of the 1st Conference on Fairness, Accountability and Transparency, in Proceedings of Machine Learning Research (Vol. 81, pp. 77–91).

Burghardt, M., Heckner, M., Kattenbeck, M., Schneidermeier, T. & Wolff, C., (2011). Design Thinking = Human-Centered Design?. In: Eibl, M. & Ritter, M. (Hrsg.), Workshop-Proceedings der Tagung Mensch & Computer 2011. uberMEDIEN|UBERmorgen. Chemnitz: Universitätsverlag Chemnitz. (S. 363-368). https://dl.gi.de/handle/20.500.12116/8018

Crenshaw, K. (1989). Demarginalisierung der Überschneidung von Rasse und Geschlecht: Eine schwarze feministische Kritik der Antidiskriminierungsdoktrin, der feministischen Theorie und der antirassistischen Politik

Criado Perez, C. (2019) Invisible Women: Data Bias in A World Designed for Men. Harry N. Abrams

Gardenswartz, L. and Rowe, A. (1998). Managing Diversity - A Complete Desk Reference and Planning Guide. New York

Haraway, D. (1988). Situated knowledges: The science question in feminism and the privilege of partial perspective. Feminist Studies, 14, 575–599. doi:10.2307/3178066

International Organization for Standardization. (2019). ISO 9241–210 Ergonomics of human-system interaction -- Part 210: Human-centred design for interactive systems, Geneva, Switzerland https://www.iso.org/standard/52075.html

Kinder, K. and Piesche, P. (2020). WAHRNEHMUNG. WAHRNEHMUNG – HALTUNG – HANDLUNG. Diskriminierungskritische Bildungsarbeit: Eine prozessorientierte Intervention. Regionale Arbeitsstellen für Bildung, Integration und Demokratie (Hrsg). https://raa-berlin.de/wp-content/uploads/2021/02/RAA-BERLIN-DO-WAHRNEHMUNG.pdf

Lee, D. (2022). Bias and Assumptions in Design Thinking. https://www.youtube.com/watch?v=UnppyWSZt2o

Link, J., Stern, P. (2021). IV. Handreichung "Ethische und sozial verträgliche KI in Unternehmen" Stand der Gestaltungsempfehlungen zu KI https://wm.baden-wuerttemberg.de/fileadmin/redaktion/m-wm/intern/Dateien_Downloads/Arbeit/Arbeitsmarktpolitik_Arbeitsschutz/Hdrg04_Stand_Gestaltungsempfehlungen_barrierefrei.pdf

Marsden, N., Link, J. and Büllesfeld, E. (2014) "Personas und stereotype Geschlechterrollen". Gender-UseIT: HCI, Usability und UX unter Gendergesichtspunkten, edited by Technik-Diversity-Chancengleichheit, Leitung Digitale Integration, Nicola Marsden and Ute Kempf, Berlin, München, Boston: De Gruyter Oldenbourg, pp. 91–104. https://doi.org/10.1515/9783110363227.91

Nielsen, J. (1994). Usability engineering. Morgan Kaufmann.

Orwat, C. (2019). Diskriminierungsrisiken durch Verwendung von Algorithmen. Antidiskriminierungsstelle des Bundes (Hrsg.) 1. Auflage. ISBN 978-3-8487-6285-9

Oudshoorn, N., Rommes, E., Stienstra, M. (2004). Configuring the User as Every-body: Gender and Design Cultures in Information and Communication Technologies, Science Technology Human Values 2004; 29; 30 DOI: 10.1177/0162243903259190

Schiebinger, L., Klinge, I., Sánchez de Madariaga, I., Paik, H. Y., Schraudner, M., and Stefanick, M. (Eds.) (2011-2021). Gendered Innovations in Science, Health & Medicine, Engineering and Environment. http://genderedinnovations.stanford.edu/methods/gender_ML.html

Wang, A., Ramaswamy, V. V. and Russakovsky, O. (2022). Towards Intersectionality in Machine Learning: Including More Identities, Handling Underrepresentation, and Performing Evaluation. FAccT '22, Seoul, Republic of Korea https://arxiv.org/pdf/2205.04610.pdf

West, M., Kraut, R., & Ei Chew, H. (2019). I'd blush if I could: closing gender divides in digital skills through education.