The Bias Network Approach: A Sociotechnical Approach to Aid AI Developers to Contextualise and Address Biases.

Gabriela Constanza Arriagada Bruneau

Submitted in accordance with the requirements for the degree of Doctor of Philosophy

The University of Leeds

School of Philosophy, Religion, and History of Science Inter-Disciplinary Applied Ethics Centre

February 2024

The candidate confirms that the work submitted is his own, except where work which has formed part of jointly authored publications has been included. The contribution of the candidate and the other authors to this work has been explicitly indicated below. The candidate confirms that appropriate credit has been given within the thesis where reference has been made to the work of others.

The work in Chapter one of the thesis has appeared in an institutional report and it is also part of a co-authored forthcoming report as follows:

1. APEC Digital Economy Steering Group. (2023). *Best Practices to Detect and Avoid Harmful Biases in Artificial Intelligence Systems* (APEC Project: DESG 05 2021A). APEC. https://www.apec.org/docs/default-source/publications/2023/9/223_desg_best-practices-to-detect-and-avoid- harmful-biases-in-artificial-intelligence-systems.pdf

2. Adasme, S., Arriagada, G., Lopez, C., & Pertuze, J. (Forthcoming). *APEC Digital Economic Steering Group (DESG), Comparative study on best practices to detect and avoid harmful biases in Artificial Intelligence systems*. Asia-Pacific Economic Cooperation (APEC)

In 2022, I joined a project with the Asia-Pacific Economic Corporation (APEC) as part of my work at the Chilean National Centre for Artificial Intelligence (CENIA). Here with other colleagues from the centre were asked to collaborate on a comparative study report about biases and bias mitigation strategies. As my PhD was on the same topic, I used some of the literature reviews I had to feed into this project, as well as use any new findings in my thesis.

In the institutional report “APEC Digital Economy Steering Group. (2023)” I helped to run the workshop to identify biases reported in the document, as well as provide the AI pipeline structure to run it. The pipeline comes from the comparative study report. During the writing of that report, my colleague Claudia López suggested we could use the same AI pipeline, and the bias and bias mitigation strategies identified there, to test the Bias Network Approach proposed in my thesis. We both contributed to building the pipeline based on the most common pipeline structures used in the AI bias literature.

In the Co-authored Comparative study report, I contributed by writing the literature review from pages 4 to 14 as lead co-author for that section with Claudia López, and with the support of Alexandra Davidoff.

A shorter version of the case study findings in Chapter Four as well as recommendations in Chapter 5 of the thesis has been published.

Arriagada-Bruneau, G., López, C. & Davidoff, A. (2025) "A Bias Network Approach (BNA) to Encourage Ethical Reflection Among AI Developers" Sci Eng Ethics 31, 1 (2025). https://doi.org/10.1007/s11948-024-00526-9

For the submitted manuscript, I drafted the main content of the article, and I provided the structure for the approach derived from my thesis work, and I am listed as the main author. We report on the core findings of the Bias Network Approach case study that are analysed in this thesis. Claudia López is the second author as she directly contributed to the design of the case study and with substantial editing and writing. Alexandra Davidoff is the third author; she contributed to the case study qualitative data analysis and editing tables and other literature review sections used.

The article is a shorter and less robust version of the theoretical background for the Bias Network Approach provided here, it is focused primarily on the case study. I collaborated with two colleagues, Claudia López (engineer) and Alexandra Davidoff (sociologist) to design and analyse the case study, as their expertise was necessary to perform this empirical portion of the thesis.

Acknowledgements

I would like to first thank my supervisor Rob Lawlor, for his outstanding support and guidance. His commitment and dedication as a professional and as a person will make me forever grateful to have been able to work with him.

I also thank Mark Gilthorpe, Peter Tennant, and Georgia Tomova, for teaching me about data analysis and causal inference, and Zach Gudmunsen and Graham Bex-Priestley, as they contributed to discussing drafts of my earlier work in the thesis as well. All of them were also a great support during the PhD.

My special thanks to my parents, Maritza and Hernán, my best friend Andy, and my husband Felipe, for their unconditional support.

I acknowledge the contributions of Claudia López and Alexandra Davidoff, who helped me implement the pilot case study in this thesis proposal. As well as the funding received by Centro Nacional de Inteligencia Artificial CENIA, FB210017, BASAL, ANID, to perform this study and the other two APEC reports referenced in this work, which allowed me to have valuable complementary literature reviews.

My doctoral studies were funded by the National Agency for Research and Development (ANID) Scholarship Program, by the Government of Chile, DOCTORADO BECAS CHILE 2018 – Folio 72190086.

Finally, I acknowledge the use of Grammarly in the proof-reading of the thesis before submission. I confirm that the proof-reading undertaken by Grammarly was in accordance with the Postgraduate Researcher Proof-reading Policy.

Abstract

In this thesis, I will introduce the Bias Network Approach (BNA) as a novel sociotechnical intervention designed to aid AI developers in identifying and addressing biases more comprehensively. The methodology of this thesis is mixed. Although primarily philosophical, it also includes an empirical case study demonstrating how to use the BNA in real life. Hence, in the second half of the thesis, I will discuss the case study findings to analyse how they support my theoretical proposal and the philosophical arguments presented in the first half of the thesis.

In Chapter 1, I will criticise “the problems of bias”: technocentrism, conceptual ambiguity, and isolated approaches to identify and mitigate bias. In Chapter 2, I will argue in favour of a sociotechnical approach, reviewing sociotechnical proposals by other researchers, and drawing key elements from them to ground my BNA proposal. In Chapter 3, I will address the conceptual ambiguity problem in more detail, contrasting “negative” and “positive” views of bias in AI, to then untangle a working definition for bias to be used in the BNA. In Chapter 4, I will present the BNA proposal through an empirical case study, analysing its main findings. In Chapter 5, I will provide guidance for developers and prompters wishing to adopt the BNA as a transitional intervention to promote ethical reflection. In Chapter 6, I will explain how responsibility should be attributed to developers, prompters, and organisations applying the BNA, including insights about responsibility as a moral obligation, forward- backward- and active responsibility, as well as the ethical agency of AI developers.

Finally, there is an Afterword, in which I will discuss an avenue for future work and hypothesise on how the core idea of a network approach could be extended to other ethical concerns in AI ethics.

Table of Contents

Introduction

Artificial Intelligence (AI) is reshaping how we live, and the more it advances, the higher the impact this technology can have in our society. Accordingly, ethical debates are a core aspect of developing AI. In the evolving landscape of AI ethics, my work will focus on one specific problem: biases. Biases in AI systems have been highly discussed because they can, among other things, perpetuate and amplify existing societal and cognitive biases. Bias in AI, therefore, demands an analysis that is not only technical but highly contextual, rooted in societal, historical, and ethical complexities.

As I will discuss in detail in the rest of this thesis, a common approach to understanding and dealing with biases in AI has been developing mitigation or "bias aware" strategies that are addressed as individual instances of bias, without engaging with the complex context of the AI ecosystem. Therefore, popular methodologies for bias mitigation tend to be predominantly reactive and centred on technocentric fixes, failing to capture the multifaceted nature of bias.

Against this idea, in this thesis, I will propose a way to reject this rather limited perspective with a more comprehensive understanding. More specifically, I will propose the Bias Network Approach (BNA), a sociotechnically inspired intervention to aid AI developers in identifying and addressing biases. With this, I will focus on aiding AI developers to engage with ethical thinking, broadening their context of analysis and, through practice, developing an awareness that allows them to think about bias by adopting a critical proactive position involving complex sociotechnical factors. Overall, with this proposal, I will aim to involve developers in an iterative, continuous process of ethical reflection, through interdisciplinary collaboration.

Consequently, the main objective of this work will be to offer a way that allows developers to engage with ethical thinking, more specifically in how they address and critically think about biases. It is important to stress, however, that the BNA does not offer a solution to all ethical issues in AI development and cannot ensure that there will be no undesirable consequences or risks after adopting this sociotechnical stance either, although associated responsibilities will be discussed in Chapter 6.

Therefore, the proposal is intended to help developers motivated to address ethical issues. The aim of applying the BNA, therefore, is to challenge developers to consider biases as integral to the design process rather than as afterthoughts, and to adopt a more robust standpoint that allows them to critically reflect about their decision- making.

To achieve this, the interdisciplinary nature of the BNA will be critical, as well as the recognition of AI systems as sociotechnical ones. AI systems do not exist in a vacuum, instead, they are part of a larger sociotechnical ecosystem. Hence, I will argue that addressing bias in AI is not just about improving how AI technology performs, but also about improving the societal outcomes and ethical standards of the technology we produce. This includes improving the professional standards and ethical insights we can demand from AI developers.

Hence, in this thesis, I will articulate the importance of the BNA as a transitional intervention to aid AI developers with their ethical reflection, I will argue for its adoption in AI development practices and discuss its potential to reshape the way AI developers think about AI bias and their related ethical choices to address them. To develop this proposal, the thesis is divided into six Chapters.

Chapter 1

Chapter 1 is divided into three main sections. In section 1.1, I will introduce the "problems of bias" which inspired my proposal for the BNA as a response to these core issues. First, I will criticise the problem of technocentrism in AI, a view that adopts an overreliance on technology to solve problems that involve societal or ethical issues, such as biases, by resorting to solving problems about technology with more technology.

Then, in section 1.2, I will present the problem of conceptual ambiguity showing that technocentric solutionism affects perceptions and actions regarding bias in AI. I will also challenge what I will call the "bias-centric views of fairness," arguing that fairness should not be defined narrowly through bias mitigation goals alone, as this affects both how fairness and bias are understood.

Finally, in section 1.3, I will identify and criticise a tendency in the AI literature I have named “the isolationist approach to bias,” consisting of treating and mitigating biases as isolated issues addressed at individual stages of AI development, which only strengthens technocentric approaches.

Chapter 1 sets the stage for the rest of the thesis, concluding with a claim for the need for AI developers to adopt a sociotechnical approach to avoid these problems, integrating social, ethical, and technical considerations to combat bias more effectively, and develop AI systems that transcend technical fixes and consider broader societal engagement.

Chapter 2

Chapter 2 is divided into four sections. After acknowledging the need for a broader context in AI ethics in the previous Chapter, in this Chapter I will argue in favour of extending the scope of inquiry for AI bias, claiming that as AI becomes more integrated into society, understanding bias requires a consideration of the interactions between technology and society, i.e., a sociotechnical approach.

In section 2.1, I will introduce the concept of sociotechnical systems drawing from sociotechnical system theory (STS), discussing its relevance to AI ethics and the importance of conceiving AI as a sociotechnical system. Then, in 2.2, I will examine the critiques of AI ethics principles by Mittelstadt (2019), Hagendorff (2020), and Munn (2023). These critiques highlight the limitations of current AI ethics guidelines and support a move toward sociotechnical approaches that better align AI principles with the practical actions of AI developers.

Afterwards, in section 2.3, I will critically evaluate Zajko’s (2021) proposal to redefine societal bias in AI, where he suggests that the term should be replaced with more robust terminology that provides a wider context for addressing societal bias. In section 2.4, I will describe a sociotechnical systemic approach to bias proposed by Draude et al. (2019), which is based on feminist epistemology and gender studies. I will draw elements from all these authors, but particularly from the ones presented in section 2.4 to develop the BNA in Chapter 4.

Overall, in this Chapter I will argue for a paradigm shift, moving from a narrow technical focus to a more integrated sociotechnical perspective that considers the ethical, societal, and technological dimensions of AI, to then use the BNA to operationalise this shift.

Chapter 3

Chapter 3 is divided into three sections.

This chapter aims to clarify how biases should be defined and understood in AI ethics, particularly in relation to the Bias Network Approach (BNA) proposal. To achieve this, Section 3.1 will introduce the three most commonly identified categories of biases in AI: technical, societal, and cognitive. While these categories are widely recognised, I will highlight the lack of focus on their interconnections and the impact of their interactions on AI development—an area where the BNA seeks to contribute.

In Section 3.2, I will examine two views on bias from a philosophical perspective, to then give a few examples from the AI literature. First, I will explore the "negative view", which conceptualises bias as morally undesirable due to its associations with unfairness, discrimination, and ethical failure. Next, I will analyse an alternative "neutral view", grounded in epistemological foundations, this perspective suggests that biases are needed to enhance efficiency in reasoning, decision-making, and problem-solving.

In Section 3.3, I will critically examine the AI examples in favour of the neutral view, arguing for a cautious approach to bias in AI ethics, highlighting further risks the authors adopting this view overlook. Accordingly, the chapter will conclude with the working definition of bias that will underpin the BNA's implementation.

This definition will emphasise the importance of carefully framing discussions around bias. I will argue in favour of adopting a negative conceptualisation of bias—at least within the specific context of the BNA's application—giving reasons to defend why it is both epistemically and ethically prudent to do so.

Chapter 4

Chapter 4 is divided into three sections. In this Chapter, I will introduce the BNA as a response to tackle the challenges reviewed in previous chapters. The BNA is characterised as a sociotechnical intervention to help AI developers contextualise and address biases by broadening the scope of analysis into their decision-making processes.

In section 4.1, I will describe the BNA, detailing its primary features and how it functions as a visualisation and mapping tool that allows developers to trace and manage the elements and factors that contribute to bias throughout the AI development process. In section 4.2, I will introduce "the waiting list project" a pilot case study to test the BNA. I will outline the qualitative methodology applied to gather insights from the development team's experience with the BNA, as well as describe the case study, which involves a retrospective examination of an NLP model development process used to identify key entities in medical and dental referrals in Chile's public hospital waiting lists.

Finally, in section 4.3, I will analyse three significant findings from the pilot case study: (i) the BNA's benefits for experimental design and revision phases, (ii) the impact of material limitations and external decisions as sources of bias, and (iii) the identification of professional biases. This analysis will be complemented by using two key concepts: "microscopic vision" from Davis (1998) and "professional deformation" from Polyakova (2014). In this Chapter, I will show how the BNA can be implemented to bridge the gap between theory and practice in AI ethics.

Chapter 5

Chapter 5 is divided into two main sections.

In this Chapter, I will outline prospective guidelines to implement the BNA within AI development projects. The chapter is divided into two main sections.

In section 5.1, I will describe how in more detail how the BNA intervention was carried out in the pilot case study, describing three stages of implementation: preliminary stage, intervention stage, and follow-up stage. Then, in section 5.2, I will highlight the benefits of the BNA for AI development.

In the chapter, I will argue that while the BNA is a structured approach, its application is flexible and can be customised to fit the particular context of each team and project. Future updates to these guidelines are expected to incorporate learnings from additional case studies, potentially offering more tailored advice.

Chapter 6

Chapter 6 is divided into four sections. In this Chapter, I will analyse the associated responsibilities of adopting the BNA as sociotechnical intervention.

In section 6.1, I will explore the concept of responsibility as a moral obligation, drawing from the work of Tollon (2022). The discussion centres on the idea that agents involved in AI creation should align their actions with societal goals and actively contribute to a better future. In section 6.2, I will distinguish between backward- and forward-looking responsibility, to then relate these types of responsibility to the BNA and the active responsibility it promotes. I will also discuss a study by Griffins et al. (2023), to complement the responsibility analysis with the ethical agency of AI developers and how this connects to some of the benefits of adopting the BNA.

In section 6.3, I will briefly comment on some of the responsibilities of different actors involved in adopting an intervention like the BNA, on an individual level (developers and prompters) and an institutional level (companies, educational institutions, and professional bodies). Finally, in section 6.4, I will conclude that the BNA offers a framework that nurtures the ethical consciousness of AI developers. I will also outline strategic reasons why even profit-oriented companies, like “Evil Corp,” might find it beneficial to adopt the BNA, such as regulatory compliance, enhancing customer trust, and systematic bias.

Conclusions, Limitations, and the Afterword

In the last section of this thesis, I will present my general conclusions and consider potential limitations to the case study and the BNA. Then, I will present an Afterword where I will discuss an avenue for future work based on the findings of the BNA pilot case study. I will use a hypothetical scenario to discuss an analogous approach called a “holistic network approach.” As this thesis unfolded, the analysis of the BNA findings showed promise in bringing ethical thinking closer to AI developers, as they were not only drawn to consider a broader context for bias assessments but also to think about ethics more generally.

This is why I see the potential to use the sociotechnical network approach in other ways, not just as a way to rethink biases in AI developmet. I will discuss how this version could help create a network of influences but for the different AI ethics principles that should be considered when tackling a new AI project.

Chapter 1: The “problems of bias”

In this chapter, I will critically examine what I call the “problems of bias.” First, in section 1.1, I will examine the pitfalls of technocentrism, an approach which supports the belief that technology inherently holds the solutions to AI problems, thus affecting the understanding we have of AI bias. I will also review different criticisms against technocentrism in the literature and argue that we should avoid this approach because it promotes a reductionist view that risks dismissing the societal, cultural, and ethical dimensions of the AI ecosystem, which ground the responsible creation and deployment of AI systems.

Afterwards, in section 1.2, I will show that in addition to technocentric solutionism influencing how developers and other people working in AI perceive and therefore address bias, there is an issue of conceptual ambiguity. This ambiguity, I will argue, promotes what I call a “bias-centric view of fairness,” that is, a narrow focus that implies defining AI fairness primarily through bias mitigation goals. To counter this, I will claim that if we want to integrate AI ethics into the work of AI developers, we need to get rid of the idea that bias is just a technical problem that can be fixed or that developers can rely solely on bias mitigation strategies to deal with biases.

Then, finally, in section 1.3, I will present a tendency I encountered during the literature review that I call an “isolationist approach to AI bias.” This isolationism entails treating biases as disconnected issues solved at individual stages of AI development, disregarding broader connections among biases and with other influential elements. I will argue that these “problems of bias” have caused critically unaware development practices in AI, promoting superficial ethical solutions. Therefore, I will call for a change in how bias is understood and approached in AI, this change consists of the integration of broader context considerations to address bias.

1.1 Technocentrism in AI

Various scholars have examined the effects of technocentrism in AI. Technocentrism can be generally characterised as an optimistic belief in the technical capabilities of AI.

Often, technocentrism is adopted by those who want to promote the advancement and development of AI, and therefore, is not presented as a conceptual stance, but rather as an adopted practice to develop the technology. Consequently, it is common to find criticisms against technocentrism based primarily on how big tech companies and AI developers adopt solutions to ethical problems, as most technocentric practices reflect data-driven objectives.[1]

It is essential to recognise that technocentric perspectives are not universal; numerous individuals or institutions that can be classified as "technocentrists" are also aware of the limitations of AI and the complex consequences of its incorporation into societal structures. The core issue with technocentrists is that they give priority to technological solutions, even in cases where the root of the problem is societal or ethical. Accordingly, a technocentrist can hold that AI, in contrast to human beings, is not limited by cognitive constraints such as biases, and susceptibilities like fatigue, stress, and social influences, all of which affect decision-making.

Hence, a technocentric view frames AI as being inherently less prone to errors that typically affect human judgment. Or it hypothesises that AI could, eventually, be freed from these human susceptibilities through the application of further technological advancements, such as the refinement of debiasing techniques within algorithms or AI models. Thus, as Peeters et al. (2021) describe it, technocentric supporters also believe in technosolutionism:

> "Although followers of techno-centrism admit that new technologies can introduce additional problems, they are also eager to point out that these problems can again be solved by applying additional technology." (Peeters et al., 2021, p. 219)

[1] I use the term data-driven here to reflect the view that considers that what drives AI development is based on the availability or capacity to analyse data, a synonym or at least related term to technocentrism. In the rest of this thesis, I will go against this interpretation of "data-driven AI," arguing to consider a broader context and operating under the assumption that it is only through human engagement that data gains significance; hence, data is not the driver of AI. Moreover, the narrative that both technologies and humans mutually shape each other is crucial, with the understanding that the significance of data is ultimately ascribed by those who develop and utilise these technologies. My stance on this issue has been influenced by Prof. Charles D. Raab, whose insights during the Leeds Data Ethics Forum at DLA Piper in late 2020 supported and complemented my reflections on the topic.

Technocentrists, based on Peeters et al. (2021) description, recognise that while AI has the potential to have a superior performance, it is not immune to the influence of human biases. Human deficiencies can carry over into AI systems through mechanisms like selection bias in the training data (Lloyd, 2018), label bias in the pre- labelling raw data (Jiang & Nachum, 2019), and inductive bias when developing AI's generalisation mechanisms (Wilson & Frank, 2023).

The point is that technocentrists believe that the solution to those ethical or societal problems lies in technical answers. For them, a perfect scenario is one in which AI functions with minimal human intervention, using its full capabilities to reduce the chances of prejudice and other errors. This viewpoint influences a conception of AI that is perceived as more impartial than human decision-making. Because of this, people turn to technosolutionism, which ultimately puts too much faith and reliance on technology to solve problems that have deeper societal roots, influenced by how people act, and how society works.

Essentially, technocentric views of AI combined with technosolutionist approaches to ethical and societal considerations, result in the oversimplification of intricate ethical and societal problems, that encompass significant issues such as prejudice, discrimination, privacy, and transparency. Technocentric oversimplification, then, stems from excessive dependence on the technology's capacity to resolve issues, which in turn implies neglecting the social factors and contexts driving AI's development.

In line with this, researchers have expressed criticisms against technocentric views that disrupt AI ethics. In what follows, I will show some examples of these criticisms, focusing on AI biases.

1.1.1 Criticisms against technocentrism

Gichoya et al. (2022), for example, present the challenge of addressing bias in medical AI algorithms, problematising that: "Due to the nature of medical data, most of the times, the classes in a dataset are very imbalanced, usually either toward the healthy prediction or the opposite." (p.32)

Additionally, the authors emphasise the high costs associated with curating and anonymising datasets, considering that complete de- identification of medical data has been proven to be unattainable (Rocher et al., 2019):

> "Labeling data are expensive, and alternatives that have been used include using medical students as labelers and Amazon Mechanical Turk workers who have no background in medicine. The cost of making datasets available is expensive, as each image is reviewed by a person to ensure anonymization. Despite several anonymization strategies, new studies show that medical data cannot be fully deidentified. This ethical conundrum will shape medical AI for decades […]" (Gichoya et al., 2022, p.563)

Aside from these general challenges, there is another specific issue the authors highlight, related to how the current evaluation of bias in medical AI is dominated by a technocentric approach, relying on statistical metrics:

> "Evaluation of bias in medical AI remains largely techno-centric, whereby statistical metrics are created for developers to evaluate if their model is biased. Statistical metrics are created for developers to evaluate if their model is biased. Examples of these metrics include true-positive rate (TPR), false- negative rate (FNR), false discovery rate (FDR), and false omission rate (FOR), among others." (Gichoya et al., 2022, p.563)

Moreover, the authors criticise that the deployment of AI in healthcare, while intended to support clinical decisions, is often driven by financial incentives, leading to a preference for AI tools, even in cases where there is evidence that AI does not outperform humans. Accordingly, the authors call for "a need to develop strategies that factor in human-machine interaction. [Because] technocentric approaches do not fully reflect the reality of medical practice." (Gichoya et al., 2023, p. 563)

For instance, they give an example of the issues that surface when technology does not align with user needs. Studies have shown, say the authors, that Electronic Medical Records (EMR) are linked to physician burnout. They claim this is increasingly integrated into healthcare as a decision-support tool, but in certain cases, AI's efficiency expectations should not be the main consideration to integrate solutions into fields like healthcare.

Matthew Bui and Safiya Noble (2020) also investigated the ethical ramifications and societal consequences of AI. In particular, the intensified scrutiny that started with prominent scandals triggered a backlash —or "techlash" as the authors call it, against tech giants such as Facebook, Amazon, Google, and other companies based in Silicon Valley. Bui and Noble argue that this criticism has been fuelled by notable events like the 2016 Cambridge Analytica scandal, which involved the improper use of Facebook data. This incident raised concerns about how AI can perpetuate bias and discrimination, as well as its ability to undermine democratic processes (i.e., the 2016 US elections and the Brexit referendum in the UK).

The rise of these criticisms has played a crucial role in bringing attention to the excessive influence of AI systems in people's lives, note Bui and Noble. It has shed light on concerns related to privacy, monopolistic control, and the inherent dangers associated with AI.

For the authors, this "techlash" has not only ignited public discussions but also raised doubts about the sufficiency of ethical frameworks to deal with these challenges. For Bui and Noble, the issue lies in the fact that tech companies' response to developing more "ethical AI" involves a technocentric approach:

> "[tech companies] seek to operationalize and create "fair" and "transparent" algorithms as a key type of intervention in an increasingly data-driven society. By and large, the emphasis on fairness interventions in AI seeks to effect and propagate technical systems that are neutral and objective and do not render any specific groups as advantaged over others." (Bui & Noble, 2020, p.165)

What the authors point out is that these efforts and responses to take an ethical stance do not emancipate AI from technocentrism. In the case of bias, more specifically, this has been a constant:

> "[The] goal of many ethics responses remains largely techno-centric, in that the goal is to perfect or "unbias" the technology, rather than account for the asymmetrical power relationships and gravity of history that renders the development and deployment of such projects deeply uneven, unethical, and even immoral." (Bui & Noble, 2020, p.166)

Thus, a critical point against technocentrism is that it promotes "unbiasing" technology as a solution, not considering the asymmetrical power relationships and historical contexts that shape technological development.

This critique leads to a broader discussion about the need for intersectional analyses to address these issues. By incorporating diverse perspectives and acknowledging the multifaceted nature of bias, we can achieve a more nuanced understanding of biases, one that goes beyond surface-level fixes and technocentric approaches, thus addressing deeper social implications.

Regine Paul (2022) also examines bias in relation to technocentric views, specifically referring to technosolutionism influencing —or delaying—AI regulation. Paul references the work of Coglianese and Lai (2022) and Krafft, Zweig, and Konig (2022) to show that technosolutionism supporters often affirm that getting rid of human bias is the main reason for using AI in public administration.

However uncritical efforts to defeat biases are criticised by researchers like Julia Powles (2018), who argues that this fixation on bias mitigation serves the big tech industry to divert attention from the need for substantial regulatory measures. Accordingly, Paul calls out this "bias obsession" arguing that there are two key elements that should be discussed to advance regulation instead of feeding this obsession:

> "a critical debate about (1) the framings of problems and solutions that get encoded in AIT regulation, and (2) the structures and power relations that render some of these interpretations policy-relevant while marginalizing others. Some research, […] duly politicizes technology regulation […]. This shows, for example, how the EU's General Data Protection Regulation's framing of 'risk' prioritizes business interests over individual rights (Padden and Öjehag-Pettersson 2021) or how the EU's emergent AI policy marginalizes questions of redistribution and economic inequality (Niklas and Dencik, 2021)." (Paul, 2022, p. 501)[2]

Paul's criticism highlights how a technocentric or technosolutionist approach to AI regulation risks overlooking its broader socio-economic consequences, such as job displacement or the concentration of economic power within specific sectors.

[2] AIT in this quote stands for "Artificial Intelligence Technology."

By prioritising technical solutions without adequately considering their societal impact, regulatory frameworks may become technically robust yet socially blind, failing to address the far-reaching implications of AI across different dimensions of society.

This critique underscores that AI regulation is not merely a technical exercise—it is inherently political, social, and ethical. The way regulatory frameworks are designed, and the narratives they reinforce, can either promote a more equitable distribution of AI's benefits or exacerbate existing power imbalances. A purely technocentric perspective on AI bias and governance tends to neglect these broader structural issues, thereby reinforcing the status quo and maintaining AI's artificial separation from the socio-political fabric that shapes and sustains it. Fjeld et al. (2020), have also noticed the problem of technocentric and limited views of bias. The authors analyse how bias is defined and mentioned across AI ethics principles and guidelines for responsible AI.

The authors examine AI principles proposed by various organisations and researchers, to identify common trends that serve as a resource for policymakers, scholars, and other actors capturing the benefits and mitigating the harms of AI. In this report, Fjeld et al. do not explicitly use the concept of technocentrism, instead, they call it a type of "potential technochauvinism." Amongst the themes identified by the authors, in the fairness and non-discrimination section they show that:

> "[…] many documents point to biased data – and the biased algorithms it generates – as the source of discrimination and unfairness in AI, but a few also recognize the role of human systems and institutions in perpetuating or preventing discriminatory or otherwise harmful impacts. Examples of language that focuses on the technical side of bias include the Ground Rules for AI conference paper "[c]ompanies should strive to avoid bias in A.I. by drawing on diverse data sets") and the Chinese White Paper on AI Standardization "we should also be wary of AI systems making ethically biased decisions". While this concern is warranted, it points toward a narrow solution, the use of unbiased datasets, which relies on the assumption that such datasets exist. Moreover, it reflects a potentially technochauvinistic orientation –the idea that technological solutions are appropriate and adequate fixes to the deeply human problem of bias and discrimination." (Fjeld et al., 2020, p. 47-48)

Here, the reference to a potential technochauvinistic orientation, is complimentary to the notion of technocentrism, although with perhaps a slight (yet relevant) difference. Based on what I have presented in this section, technocentrism refers to the belief that technology is the central or most important approach to solving problems in AI. Therefore, technocentrism involves the assumption that technological progress is inherently positive and that any challenges or issues can be solved with more or better technology. This can lead to an overreliance on technology without sufficient consideration of the social, cultural, or ethical factors that are also at play.

Technochauvinism builds on this. The term is often attributed to Meredith Broussard, who defines it in her book "Artificial Unintelligence: How Computers Misunderstand the World" (Broussard, 2018). Broussard claims that overconfidence in AI's ability can be developed to the point of chauvinism —a belief that technology is always the best solution, leading to dismissing or devaluing human skills, and expertise, as well as other valuable traditional methods.

The key difference between the two is that while technocentrism places technology as the desirable solution, technochauvinism goes beyond asserting the superiority of technology, in a way that overrides the importance of human capabilities and non-technological solutions.

Broussard comments on how to confront this chauvinism, saying that we should find technological solutions by questioning which instrument is better suited for the job (Broussard, 2019). This means that in certain cases, AI can be the most suitable instrument to achieve a specific outcome, however, on other occasions, the simplicity and value of human expertise might be the best option available. Thus, for Broussard, to counteract the bias of technochauvinism, an approach that combines the strengths of humans with machine capabilities is fundamental. Nevertheless, this collaboration is not easy to achieve, especially when bias-related problems continue to surface.

1.1.2 Avoiding technocentric views.

Technocentrism, as discussed above, favours technological solutions at the expense of considering social, cultural, and human factors, which I believe is an undesirable approach to conceptualising and dealing with ethical issues in AI, particularly biases—

hence why my work will involve adopting a sociotechnical view of AI (see Chapter 2) in response to technocentrism.

Technocentrism can also influence how developers perceive and address bias in AI systems. By prioritising technology as the central solution, technocentrism can lead developers to view bias as a technical problem that can be resolved with more technology, such as advanced bias mitigation techniques. This perspective tends to oversimplify the complex nature of bias, which is not merely a by-product of flawed technology but also a reflection of deeper societal, cultural, and historical prejudices that are embedded in a model's training data and the design choices developers face.

Consequently, under a technocentric view, developers might overlook the critical need for a multi- and interdisciplinary approach that includes ethical considerations, sociological insights, and overall perspectives beyond the specific biases they wish to address. The technocentric lens can narrow their vision to quantifiable aspects of bias, potentially neglecting the subtler, qualitative dimensions of fairness and discrimination that are closely related to bias. This can result in AI systems that perpetuate or even exacerbate existing inequalities, as they are designed and refined within an echo chamber of technological optimisation where technological failures are fixed with more technology, disconnected from the broader context in which they operate.

Thus, in the next chapters, I will argue that embracing a broader context is crucial for integrating AI ethics into the fabric of developers' professional practice, helping them move beyond technocentric views that can lead them to resort to technical solutions as a default answer to solve complex issues about bias.

But, before examining the sociotechnical aspects contributing to countering technocentrism, I will review two other "problems of bias" derived and influenced by technocentrism: (i) the bias-centric view of fairness and (ii) the isolationist approach to biases. In the next section, I will explain and examine (i), showing that the technocentric narrative is present not only in how organisations and big tech shape the discourse but also in crucial practices of AI developers.

1.2 Conceptual ambiguity and the bias-centric view of AI fairness.

In addition to technocentric solutionism influencing how developers and other people working in AI perceive and therefore act on bias, there is another issue we face: conceptual ambiguity. Witthlestone et al. (2019) in their report on "Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research" claim that a crucial obstacle, to solving ethical and societal issues raised by algorithms, data, and AI is:

> "[…] the ambiguity of many central concepts currently used to identify salient issues […] like 'fairness', 'transparency' and 'privacy' […]. While they have served to highlight common themes emerging from case studies, many of these terms are overlapping and ambiguous. This stems partly from the fact that different fields, disciplines, sectors, and cultures can use these concepts in substantially different ways, and partly from inherent complexities in the concepts themselves. As a result, discussions of the ethical and societal impacts of ADA risk being hampered by different people talking past each other." (Whittlestone et al., 2019, p.14)

Stressing this, I hypothesise that a contributing factor to prioritising technical solutions to deal with bias is precisely rooted in these conceptual challenges of bias.

1.2.1 Conceptual ambiguity: a conflation of "bias" and "unfairness".

As Whittlestone et al. point out there is an ambiguity surrounding bias, more specifically a conceptual ambiguity that stems from the multiple meanings, uses, and aspects of bias studied across disciplines.

As an example, they contrast the use of "biased sample" in statistics —simply meaning an inadequate representation of the feature distribution in the reference population— and the use of bias in psychology or law, where bias carries a normative negativity demonstrating a prejudice towards specific groups or individuals. This can create an issue, say Whittlestone et al., because:

> "a dataset which is 'unbiased' (in the statistical sense) may nonetheless encode common biases (in the social sense) towards certain individuals or social groups. [Therefore], distinguishing these different uses of the same term is important to avoid cross-talk." (Whittlestone et al., 2019, p. 15)

More clarity regarding these distinctions will be discussed in Chapter 3. Also, to avoid "crosstalk," related concepts like fairness require attention, say the authors.

Highlighting how different disciplines have their own research cultures, that influence the way they conceive a complex concept like bias or fairness. For example, they mention that in machine learning (ML), researchers are naturally driven to construct a mathematical definition of fairness, whereas in other areas, the approach to the concept changes:

> "[Q]ualitative social scientists would often seek to highlight the rich differences in how different stakeholders understand the concept. Similarly, philosophical ethicists often seek to highlight inherent dilemmas and in-principle problems for different definitions of a concept, whereas many lawyers and researchers from other policy-oriented disciplines would look for operational definitions that are good enough to resolve in-practice problems." (Whittlestone et al., 2019, p. 15)

This becomes even more challenging when terminological overlap happens. The authors define this as the case when different terms are used to "express overlapping (though not necessarily identical) phenomena," (Whittlestone et al., 2019, p. 14) exemplifying that this is common in AI where terms like "bias", "fairness", and "discrimination" refer to some type of disadvantage for individuals or groups caused by problems in datasets or algorithmic models.

The authors describe this conceptual ambiguity by stating that "the terms 'bias' and 'fairness' are often conflated, with some discussions of such cases simply defining bias as unfair discrimination." (Whittlestone et al., 2019, p.16) They show that crosstalk often happens, and that, in practice, there is a tendency to conflate these terms as they are assimilated as referring to one phenomenon, instead of different interacting phenomena. On this point, however, I have to make some distinctions.

Although I agree with what the authors are conveying with this conceptual ambiguity, that is, a type of conflation between bias and fairness, it would be more appropriate to say that the terms 'bias' and 'unfairness' are the ones often conflated. What is commonly associated with bias is either a technical or human error that results in a type of prejudice that affects individuals and groups arbitrarily— without a relevant or justified reason to support it. The conceptual ambiguity highlighted by Whittlestone et al., however, shows another problem I will call the "bias-centric view of fairness."

1.2.2 A bias-centric view of fairness.

With a bias-centric view of fairness, I refer to a narrow focus on the identification and mitigation of biases as a response to fairness concerns, implying that their elimination leads to achieving fairness —thus, prioritising a formalised (technical) definition of fairness as well as a technical mitigation strategy to address biases.

I will argue, however, that this view is overly simplistic because it fails to recognise the full spectrum of what constitutes fairness, which is not merely about removing bias but also about ensuring equity, justice, and inclusivity. And, simultaneously, it contributes to limiting the role of bias, to be dependent on what fairness definition operates in AI development, which translates into an understanding of bias as something "technically fixable." Hence, concentrating solely on bias in relation to fairness or by overlapping the terminology of "bias" and "unfairness" i.e., a bias-centric view of fairness, risks overlooking the contextual influences both concepts have in AI development.

An example of the risks related to ignoring these contextual influences is discussed by Jamie Brandon (2021), who warns that debiasing efforts could inadvertently eliminate useful biases, thus hindering a comprehensive understanding of fairness. This myopic stance on bias can obscure rather than clarify broader ethical issues, preventing developers from effectively tackling the nuanced complexities of fairness in AI systems. Brandon explains that:

> "Debiasing a model censors out insights around unethical bias. Analysts need to be able to find these insights and share them with management, peers, and their society. Evidence shows transparency with biases instigates change. By choosing not to debias the model, the analyst will be more aware of problems that exist. Just as reporting the wage gap in salaries, the model showing its flaws can lead to solutions in society rather than just in the model. Before finding a solution, awareness and a proper understanding of the problem are key." (Brandon, 2021, p.107)

Brandon points out the importance of guiding the AI developer in handling ethical biases. For the author, leveraging the model's unethical results helps create awareness. Therefore, "blinding the analyst to the issue at hand doesn't solve the problem." (Brandon, 2021, p.107) In contrast, acknowledging and understanding biases as part of the context they need to face, aids developers in building more robust evaluations of their practice, as well as achieving further insights about how biases are influencing the AI development pipeline.

Thus, I consider that the bias-centric view of fairness often stems from a technocentric mindset (although sometimes unconscious) that seeks technical solutions to complex societal challenges. Under this mindset, biases are measured and quantified as indicators of fairness, promoting the illusion that technical fixes are sufficient to address biases in AI, or that by mitigating biases, developers are making enough efforts to develop more ethical AI systems.

A bias-centric view of fairness, therefore, fails to consider that fairness and bias are concepts that transcend technical dimensions and require broader societal engagement and that they should also be conceptually independent. AI bias, for instance, should prompt a deeper inquiry into what that bias is telling us about systemic issues (e.g., deeper social injustices) and not just be viewed as a problem to be solved through algorithmic or statistical adjustments. Furthermore, as I will discuss in my proposal of the Bias Network Approach in Chapter 4, biases that do not lead to actionable harm can still reveal aspects of the societal context. It is a comprehensive view of biases, I will argue, the one that could lead to a better understanding for AI developers to make sound ethical decisions.

Accordingly, adherence to technocentrism and a bias-centric view of fairness can result in what Ben Green and Salomé Viljoen (2020) describe as a belief in the neutrality of computational solutions, often ignoring that technocentric solutions can perpetuate existing power structures and more profound societal issues. The emphasis on quantifiable solutions as a response to normative issues usually equates objectivity with value neutrality. However, this supposed neutrality often mirrors the prevailing values of dominant social groups, essentially masking their perspectives as the norm:

> "The algorithmic formalist emphasis on objectivity and neutrality occurs on two related levels. First, algorithms are perceived as neutral tools and are often argued for on the grounds that they are capable of making "objective" and "neutral" decisions. Second, computer scientists are seen by themselves and others as neutral actors following the scientific principles of algorithm design from positions of objectivity." (Green & Viljoen, 2020, p.21)

I will analyse and respond to these "objectivities" in Chapter 2, where I will discuss the importance of seeing biases not only from a sociotechnical view but from the situated knowledges (Harding, 1992, 2015) of AI developers, as well as under the critical notion of strong objectivity (Haraway, 1988). And in Chapter 3, I will also comment on a study about the perceptions of AI developers about their ethical agency and the neutrality of AI.

Thus, I will argue in further chapters that to achieve a meaningful integration of AI ethics into developers' practice, it is necessary to reject the idea that bias can simply be solved with technical tools, and that bias mitigation strategies are sufficient for developers to address biases.

Therefore, instead, we must hold AI developers accountable for their choices and the potential harm these choices can cause, as emphasised by Hooker (2021). Hooker presents a nuanced perspective on algorithmic bias, emphasising that it is crucial to understand the sources of bias, as this knowledge informs us to identify the efforts that are the most appropriate to mitigate harm. If bias was merely a data issue, the best solution would be to adjust the data handling processes. However, says Hooker:

> "data ''fixes'' such as re-sampling or re-weighting the training distribution are costly and hinge on (1) knowing a priori what sensitive features are responsible for the undesirable bias and (2) having comprehensive labels for protected attributes and all proxy variables." (Hooker, 2021, p.1)

This condition of satisfying (1) and (2), says Hooker, is often unattainable. With the complexity of real-world data, particularly in fields dealing with images, language, and video, labelling every pertinent feature is an unrealistic goal, says the author.

Moreover, even if we were to label sensitive attributes like race and gender adequately, algorithms might still identify and utilise proxy variables to infer these attributes. Hooker considers that this challenge extends to the collection of protected attributes, which can be burdensome and further complicated by the lack of consistent categorisation across datasets.

Given these obstacles, Hooker points out that the overall harm within a system is not solely a consequence of the data but also the result of how models are designed and developed —this is the key. Recognising the impact of model development on potential harm is critical because some design decisions can lead to better outcomes than others. Hooker's argument stresses that a comprehensive understanding of both the data and model design is essential to reducing harm. This recognition, however, needs to be paired with a change in how bias mitigation is approached. Accordingly, there is another "problem of bias" that needs attention to overcome the technocentric views of AI and bias, and the bias-centric views of fairness. This is the isolationist approach to AI bias.

1.3 The isolationist approach to AI bias.

One final problem of bias I will discuss here is the isolationist approach to addressing bias in AI. So far, I have highlighted different technocentric criticisms that have been raised by other researchers, as well as highlighting the conceptual ambiguity in how bias and fairness are connected, and the prevailing bias-centric view of fairness, where the presence or absence of bias determines fairness definitions and goals. Here, however, I will talk about an issue that I have identified when performing a literature review on bias mitigation strategies: the isolationist approach to bias.

The isolationist[3] approach to bias refers to the tendency in the literature where biases are attributed to specific steps in the AI process, leading to mitigation strategies that are limited to addressing individual stages within the AI pipeline to prevent concrete harm or effects. But how did I come to distinguish this approach?

[3] Before naming this the "isolationist approach", I had just described the phenomenon and presented these findings in a work-in-progress session. In this session, Alfonso Donoso suggested the name "isolationist" to describe it. I am grateful to his contribution and feedback during that session.

In related co-authored work (Adasme et al., Forthcoming; APEC Digital Economy Steering Group, 2023) I wrote two literature reviews that gave me further insight into this issue. One literature review was used to build the AI pipeline in Figure 1:[4]

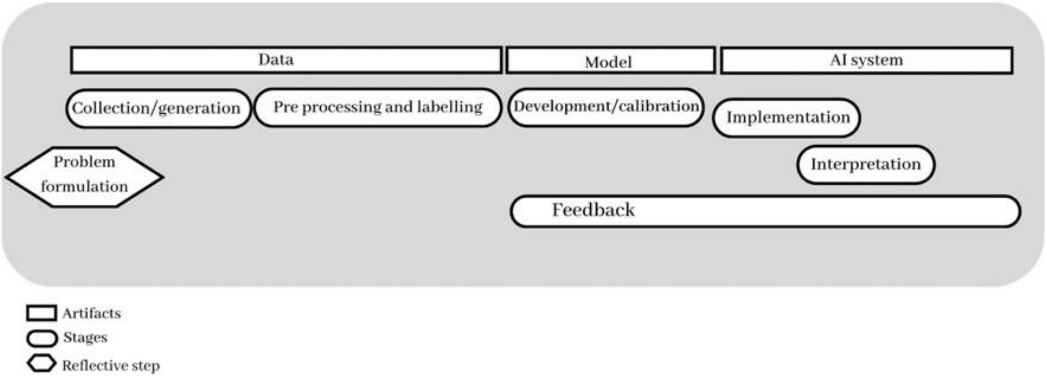


Figure 1: Basic structure of the AI pipeline.

The pipeline is based on the most common steps highlighted in the AI bias literature and recurrently used in studies about bias mitigation in AI, as these explicitly describe how biases are located in the pipeline. The specific papers consulted in this review to build the AI pipeline can be found in Table 1 in the Annex.

The structure of this AI pipeline was constructed based on commonly highlighted steps in the literature. It starts with the problem definition, followed by data collection or generation, and tasks involving pre-processing (such as data cleaning) and labelling, which are, in turn, linked with data artefacts. Subsequently, the model artefact is related to model development, involving activities like model training, calibration, performance comparison, and model selection. Finally, this model is implemented and integrated into an AI system, whose outputs users can interpret.

Notably, feedback occurs at various stages, including during model development and post-deployment of the AI system. In the literature reviews it was possible to see how most of the biases documented in prior literature are associated with one or just a few procedural stages, as shown in Figure 2 —which includes a couple of examples.

[4] I use this pipeline to analyse the bias mitigation strategies I will discuss in this Chapter, and for the BNA proposal in Chapter 4.

While it is possible to recognise that different biases can cover broader ranges of the pipeline, they are generally focused on sub-parts of it. For example, societal biases are often mentioned at the beginning of the AI pipeline (Suresh & Guttag, 2021). These biases are commonly identified from problem formulation to data-related steps or data encoding patterns of historical discrimination shown in the quality of the data and representation bias (Char et al., 2020; Dobbe et al., 2018).

In the case of cognitive biases, they are mainly associated with algorithmic model development affecting its design, for example in variable analysis (Srinivasan & Chander, 2021), calibration and evaluation (Char et al., 2020; Olteanu et al., 2019) and improper variable use and skewed data (Sangokoya, 2020, Fazelpour & Danks, 2021). Regarding technical biases, various researchers highlight data collection issues related to sampling, measurement, and selection biases (Akter, McCarthy, et al., 2021; Cramer et al., 2018; Srinivasan & Chander, 2021). As well as processing biases which hinder models' learning and generalisation, like aggregation biases.

Overall, the literature reviews unveiled a trend of addressing biases individually, or as I call it, in an isolationist way, which in turn promotes the development of targeted tools to address specific types of biases and their effects on system performance, i.e. isolationist mitigation strategies. Hence, I hypothesise that researchers' focus on creating and implementing mitigation strategies as the primary response to unwanted biases leads to an isolationist perspective. This perspective tends to conceive mitigation strategies as targeted fixes, missing the broader societal and structural aspects of biases that shape the AI development process —a recurrent critique present in the technocentric and the bias-centric view of fairness criticisms above.

Now, I do not want to be misunderstood. My constant criticism against technocentric views and the recurrent priority given to formalised technical solutions does not imply that I consider these advances in fairness formalisations or bias mitigation tools as undesirable. On the contrary, these strategies are crucial and have proven effective in dealing with specific issues. For instance, to combat racial bias due to underrepresentation in data sets, methods like oversampling have been used to improve inclusivity.

Oversampling is a technique used to alter the distribution of classes in a dataset. It is especially helpful in situations where there is an unequal distribution across classes, with one class being significantly larger than the others. The main goal of this technique is to increase the occurrence of the minority class in the dataset to achieve a more equitable distribution, guaranteeing that each class is adequately represented in the data.The work of Buolamwini and Gebru (2018), for example, shows that incorporating more diverse skin tones into data sets, enhanced facial recognition accuracy. Likewise, advanced techniques like the Synthetic Minority Oversampling Technique (SMOTE) have been developed to specifically reduce biases associated with labels and selection (Zhou et al., 2023).

Additionally, one of the latest approaches applies concepts from differential privacy to handle bias as if it were data leakage during training. Thus, not requiring prior knowledge about potential biases in the data, such as those found in word embeddings, relying on unsupervised methods to foster fairness (Liao et al., 2023). In the literature reviews, several mitigation strategies were also identified —see Figure 3 for some examples— but just as in the case of biases (Figure 2), these strategies are mentioned, defined, and analysed as isolated solutions for specific biases or stages in the AI pipeline.

In the initial stages of the AI pipeline, the identification of sensitive subjects or attributes (Feuerriegel et al., 2020; Srinivasan & Chander, 2021) was a prevalent type of mitigation, promoting the incorporation of anti-discrimination measures. Bias mitigation strategies usually applied during data collection include scrutinising existing datasets, especially canned datasets (Olteanu et al., 2019). Bias awareness and high-quality standards checking source selection, methodologies, instruments, and labelling (Baker & Hawn, 2021; Ntoutsi et al., 2020) are also common practices.

During the stage of algorithmic modelling, techniques such as adversarial debiasing and the reduction of disparate impact are often implemented, as noted by Rovatsos et al. (2019).

However, there is a scarcity of such strategies in the later stages of the AI pipeline. These involve evaluating how the model's performance aligns with the project objectives, conducting internal assessments, and ensuring predictions are generalisable across different subpopulations (Baker & Hawn, 2021).

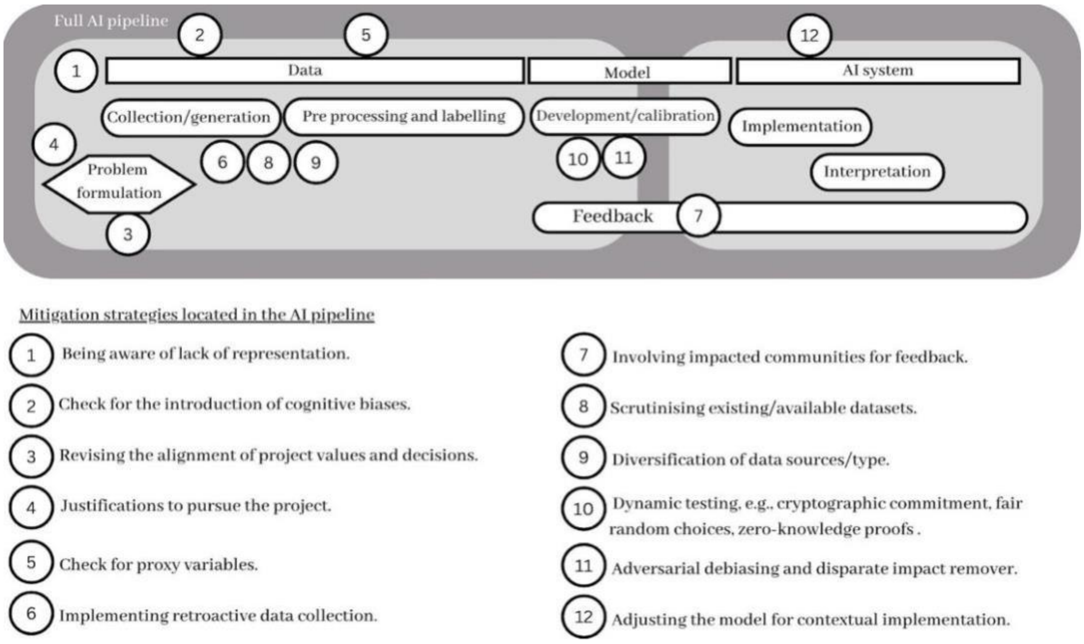

Full AI pipeline
Data
Model
AI system
Collection/generation
Pre processing and labelling
Development/calibration
Implementation
Interpretation
Problem formulation
Feedback
1 2 3 4 5 6 7 8 9 10 11 12
Mitigation strategies located in the AI pipeline
1 Being aware of lack of representation.
2 Check for the introduction of cognitive biases.
3 Revising the alignment of project values and decisions.
4 Justifications to pursue the project.
5 Check for proxy variables.
6 Implementing retroactive data collection.
7 Involving impacted communities for feedback.
8 Scrutinising existing/available datasets.
9 Diversification of data sources/type.
10 Dynamic testing, e.g., cryptographic commitment, fair random choices, zero-knowledge proofs .
11 Adversarial debiasing and disparate impact remover.
12 Adjusting the model for contextual implementation.

Figure 2: Common bias mitigation strategies prevailing in the literature

Bias mitigation techniques in later stages also include improving the interpretability of models (Kizilcec & Lee, 2022) by involving impacted communities and making the scrutiny of datasets more accessible, particularly common during implementation and feedback stages.

My criticism here, once again, is that these methods of addressing bias tend to isolate specific biases and recommend tools or actions to mitigate their effects on system performance, without requiring or stimulating a critical evaluation or consideration for a broader context. Accordingly, if there is any contextualisation, it is generally limited to the technical effects of biases, rather than a comprehensive examination of how various aspects of AI development might contribute to biases throughout the pipeline.

Therefore, I argue that while mitigation strategies are valuable, they are often confined to discrete efforts and limited to specific stages of the AI development pipeline, encouraging an isolated and de-contextualised understanding of bias.

Hence, although these mitigation methods are important components of AI development, they should not be the sole or primary means by which developers comprehend and address biases —as I will argue in the rest of this thesis.

In summary, I criticised the isolationist approach to bias mitigation, stressing that while it can be methodically sound, to overcome the problems of bias outlined here, developers need a more systemic and integrated perspective to deal with biases.

1.4 AI development needs to look at a broader context.

In this first chapter, I examined the "problems of bias" criticising the prevalent technosolutionist approaches, more specifically the conceptual ambiguities and conflations of bias and unfairness, that contribute to a bias-centric view of fairness and favour isolated strategies for mitigating biases.

While I do not question the importance of reducing bias in certain contexts, I do argue that choosing the most effective strategy for addressing undesirable biases demands more than just technical fixes and procedural measures.

I claim that a robust ethical approach for addressing biases in AI necessitates the incorporation of critical ethical deliberation, especially within AI developers' decision-making —to actively avoid the problems of bias examined here. To cultivate this ethical reflective stance in AI developers, I will propose a sociotechnical approach as a response to the technocentric tendencies that I have scrutinised here.

With this shift into a sociotechnical approach, I will argue that the AI bias debate needs a re-evaluation of how bias problems are defined and acknowledged by AI developers. Therefore, the reasoning behind selecting bias mitigation strategies will extend beyond the narrow focus on individual biases and their immediate impacts.

My proposal will be based on the belief that to effectively resolve issues about biases, AI developers must extend their perspective to broader contextual factors. This proposal is my direct response to the criticism against the isolationist approach and its limited view, which isolates the perception of AI biases as individual problems rather than recognising them as interconnected elements rooted in larger societal and ethical issues.

The issue is that if developers only associate biases with particular stages of development like data collection, model training, or deployment, they are at risk of adopting short-sighted strategies that only fix immediate technical problems. This results in them neglecting the bigger picture of how these stages are interconnected and influenced by societal contexts, making AI developers prone to falling into these "problems of bias."

This limitation can prevent AI developers from engaging in a deeper ethical discourse, fostering a simplistic, solutionist mindset instead of promoting an in-depth understanding of biases. Hence, "isolationist mindsets" fail to recognise how addressing biases is, essentially, a sociotechnical concern, dismissing the interaction that biases can have even within the AI pipeline.

For example, societal biases that skew problem definitions in earlier stages can lead to biased data gathering, which then affects model training and deployment. Each phase is built on its predecessor and failing to address the possibilities for this network of influences, can exacerbate biases throughout the development process.

Therefore, my critique against technocentrism, the narrow bias-centric views of fairness and the isolationist approaches is a call to action for rethinking how AI developers understand and address AI biases. Fundamentally, my critique demands a new way of addressing AI bias, integrating social, ethical, and technical considerations into a unified approach.

Such an approach, I will argue, would not just combat biases more effectively; it can also aid AI developers to reflect on the broader implications of their work, thus preventing them from engaging in “solutionist ethics.”[5]

Consequently, in the following chapters (more specifically chapters 3, 4, and 5) I will suggest that if AI developers want to adopt an effective bias-aware perspective, they must thoroughly examine biases and their origins as part of a unified process, i.e., a network approach. The objective of the approach will be to counteract the inclination towards isolationism by highlighting the interconnectedness among biases in AI, amongst themselves, and with external influences.

Thus, following West et al. (2019), I will address the need to broaden the scope of discussion about AI bias:

> “As the focus on AI bias and ethics grows, the scope of inquiry should expand to consider not only how AI tools can be biased technically, but how they are shaped by the environments in which they are built and the people that build them.”(emphasis added) (West et al., 2019, p.6)

In the following chapter, I will show how, while the case for a broader context in AI ethics has been argued before, it remains primarily theoretical. By critically examining various sociotechnical approaches and explaining why I choose to anchor my approach inside this framework, I will lay the groundwork to introduce my distinctions to define bias in Chapter 3 and present the BNA proposal in Chapter 4.

[5] Solutionist ethics, as defined by Evgeny Morozov (2014) and further elaborated on by Oliver Nachtwey and Timo Seidl, (2023) is an ideology that interprets complex social situations as problems that can be neatly defined and solved using computational solutions or algorithms —quite similar to the vices in technocentric approaches. Solutionist ethics posits that virtually social issues can, in principle, be addressed technologically, akin to having a technological “hammer” for a social “nail”. Solutionist ethics, hence, is a view suggesting that social problems are not necessarily rooted in power or wealth asymmetries requiring political solutions, but rather in inefficiencies and deficiencies that can be rectified through appropriate technological interventions. In essence, solutionist ethics values the use of technology as the primary means to resolve social challenges, side-lining other considerations like political or socioeconomic factors.

Chapter 2: The Need for Broader Context in AI Ethics: Adopting a Sociotechnical Approach

In the preceding chapter, I introduced the need for a sociotechnical approach, departing from the isolationist and technocentric tendencies I criticised and following West et al.'s (2019) call to extend the scope of inquiry for AI bias. This shift, as I will argue in Chapter 4, is more than just a change in focus; it is a change in the perspective that developers must adopt to deal with biases.

As AI technologies continue to be embedded into the social fabric, it becomes evident that any analysis of bias must take into consideration this intricate interaction between technology and society. The following discussion in this chapter encourages us to look beyond the technical elements of AI and consider a broader ethical context.

Accordingly, in this chapter, I will illustrate sociotechnical approaches, highlighting that they are predominantly theoretical. Hence, I take some insights from these contributions to construct my Bias Network Approach, emphasising the gap between theoretical sociotechnical frameworks and their practical application.

I will start section 2.1 by giving a background to the notion of sociotechnical systems and the adoption of this concept in AI ethics. Then, in section 2.2. I will examine three proposals by Mittelstadt (2019), Hagendorff (2020), and Munn (2023) criticising AI ethics principles. All these authors, although not explicitly, adopt a sociotechnical view or promote their foundational elements. They share a set of commonalities in their criticisms of the limitations of AI ethics guidelines and suggest we improve the "undesirable status quo"—a proposal that is consistent with my criticism against the isolationist and technocentric views presented in Chapter 1.

These views support the need for a shift to sociotechnical approaches that engage with the practices that AI developers can adopt, given that AI ethics has long been criticised for having a gap between principles and practice. With that gap in mind, I will examine specific sociotechnical conceptualisations of AI bias.

First, in section 2.3, I will evaluate the benefits and limitations of Zajko's (2021) redefinition of societal bias, where he argues that the notion should be replaced by more robust terminology, supporting a broader context to discuss and deal with societal bias in AI.

Then, in section 2.4, I will describe Draude et al.'s (2019) sociotechnical systemic approach to bias, grounded on feminist epistemology and gender studies, from which I will draw various elements that I will use as a conceptual basis for my BNA proposal in Chapter 4.

2.1 AI as a sociotechnical system.

The sociotechnical approaches in AI ethics find their origin in Sociotechnical Systems Theory (STS), which considers the development of sociotechnical systems as a dynamic journey rather than a static endpoint.

Chris Clegg (2000) emphasises that design is a continuous process extending beyond initial implementation. Users perpetually reinterpret and adapt the system in real-world scenarios; hence the system's evolution is shaped by its contextual use. Sociotechnical theory, thus, underlines the significance of social variables, asserting that changes in organisational systems are partially propelled by the social dynamics at play.

Clegg also claims that within the framework of STS, when it comes to weighing societal and technical aspects and principles "none should take logical precedence over the other, and that they should be designed jointly." (Clegg, 2000, p.465) This is particularly salient in AI development, where developers must anticipate and mitigate the potential effects of AI systems, effects that are often elusive at the outset.

For instance, unforeseen consequences of AI systems are more commonly spotted post-deployment, revealing latent social and ethical risks. Therefore, a sociotechnical approach generally requires a holistic assessment that spans both the pre-deployment and post-deployment phases of a complex system, pursuing an iterative and reflective process to address the extensive societal implications of AI.

Sociotechnical insights can frame how we understand the AI ecosystem. When we venture into the development of AI, we are navigating a complex ecosystem of values, decisions, resources, and processes that need to be carefully considered and balanced. Now, translating the general insights from STS into AI has been a relatively straightforward shift.

Researchers who call to adopt a sociotechnical view of AI, fundamentally criticise the idea of isolating technology from the organisational and social environment in which it is created. These perspectives shed light on the complex relationship between technological innovations and the societal constructs they interact withas Niehaus and Wiesche (2021) have articulated.

In the case of AI, sociotechnical approaches provide a framework to analyse the dynamic interplay among the different actors and elements that constitute the broader societal context that informs the development of AI.

Loi et al. (2021), for example, incorporate a sociotechnical view to analyse design explanations in algorithms. Benk et al. (2022), adopt this view for measuring trust, informing human-AI interactions. And van de Poel (2020), defines AI systems as sociotechnical systems. Van de Poel aligns them with the traditional definition of sociotechnical systems involving a combination of technical components, human agents, and institutional elements, with an account for determining when AI systems can embody certain values.

Therefore, what makes AI systems distinct in a sociotechnical context, as highlighted by van de Poel, is their unique ability: "[An AI system can] autonomously interact with its environment and adapt itself based on such interactions." (van de Poel, 2020, p. 387) And this interaction is, unavoidably, rooted in the context in which AI is designed and used.

Hence, adopting a sociotechnical view of AI implies establishing a broader context of analysis, and understanding how AI systems are embedded in a sociotechnical context that frames their development, and that is influenced by their interaction with this broader context.

2.2 When AI principles and guidelines are not enough.

Within this discussion of the need for a broader context in AI ethics, there is a particular debate about the efficacy of AI ethics principles that exposes a crucial gap between theoretical frameworks and real-world applications. However, this transition from abstract principles to actionable guidance is not straightforward; it requires a deep understanding of the sociotechnical landscape in which AI operates.

And, since my goal is to contribute with an approach to support AI developers in their ethical reflection to address biases through a networked understanding of them, it is important to gather initial insights to develop this sociotechnical intervention. By examining this debate, I will argue, that it becomes apparent that the successful implementation of ethical principles in AI requires an ecosystem approach, one that accounts for the interplay of technology, society, and human behaviour. One to which the BNA will contribute.

2.2.1 Ethics as a process, not a destination.

In 2019, Brent Mittelstadt brought attention to the issue of depending on ethical principles to guarantee ethical AI development, as this has significant limitations. Mittelstadt, after analysing several public-private initiatives defining values, principles, and frameworks for ethical AI, argued that rather than offering concrete, targeted recommendations, "many initiatives, particularly those sponsored by industry, have been characterised as mere virtue signalling intended to delay regulation and pre-emptively focus debate on abstract problems and technical solutions." (Mittelstadt, 2019, p.501)

As a result of this phenomenon, says Mittelstadt, ethical standards are viewed as abstract high-level principles and value declarations that ignore the basic normative and political conflicts present in important AI concepts such as privacy and fairness.

Mittelstadt's analysis contrasts AI with other fields such as medicine, where ethical practice is an established standard. In AI ethics the goal of incorporating ethical principles into professional practices or providing a framework to improve the development of AI systems raises concerns. Given the lack of professional traditions and structures that can facilitate this transition, some of these concerns are about how these principles can be translated into the AI context.

To analyse these concerns, Mittelstadt makes a comparison between the domains of medicine and AI. In medicine, there is a Hippocratic tradition, a professional ethics with a clear set of ethical standards, for example, to aid in the identification of morally questionable clinical treatments or human trials. These considerations, which focus on both conduct and practice, are central to medical professional training and policymaking.

This historical regulatory influence has a significant impact on the ethics of medical practitioners and medical institutions, which has yet to be replicated in the field of AI —and may be more difficult to achieve given the influences and incentives driving AI development in the private sector.

In medicine, for example, there is a clear overarching goal of promoting the well-being of patients: "It is a defining quality of a profession for its practitioners to be part of a 'moral community' with common aims, values, and training" (Mittelstadt, 2019, p.502). This common aim provides a collective understanding and application of ethical principles, as well as their translation to professional codes and standards of practice.

Health professionals have fiduciary duties to their patients, says Mittelstadt, with established ethical obligations. In AI, the goals are more diverse and often commercially driven, making it harder to apply a unified ethical framework:

> "AI development is not a formal profession. Equivalent fiduciary relationships and complementary governance mechanisms do not exist for private sector AI developers. AI developers do not commit to 'public service', which in other professions requires practitioners to uphold public interests in the face of competing business or managerial interests." (Mittelstadt, 2019, p.503)

Furthermore, the medical field is regulated by robust legal and professional accountability mechanisms, such as malpractice laws and ethical committees, ensuring adherence to ethical standards. AI, however, lacks such comprehensive accountability mechanisms, making it harder to ensure ethical compliance:

> “Excluding certain types of risks (e.g. privacy violations governed by data protection law), AI development does not have comparable professionally or legally endorsed accountability mechanisms. […] Long-term commitment to self-regulatory frameworks cannot be taken for granted. Prior research on the impact of codes of ethics on professional behaviour has revealed mixed results. Codes are often followed in letter rather than spirit, or as a ‘checklist’ rather than as part of a critical reflexive practice.” (Mittelstadt, 2019, p.507- 8)

As a response to this challenge —i.e. overcoming the gap between principles and ethical practice— Mittelstadt suggests at least four ways to address the insufficiency of principles to guarantee ethical AI.

First, he suggests that AI Ethics initiatives should clearly define their long-term aims and impact, with binding accountability structures at organisational levels that include “cooperative oversight to ensure translated norms and requirements remain fit for purpose and impactful over time.” (Mittelstadt, 2019, p.509) This would involve establishing professional and institutional norms through inclusive design, transparent ethical reviews, documentation, and independent ethical auditing. These steps are presented as necessary to ensure that the norms and requirements are defining sustainable pathways for the future impact of AI.

Mittelstadt also encourages a ‘bottom-up’ approach to AI ethics in the private sector. This approach would require recognising the diversity in AI technologies, including both generalist ‘top-down’ and localised ‘bottom-up’ approaches to AI systems, involving collaborative assessments to specify principles and set precedents, moving professional standards forward.

The focus, accordingly, should be on developing sector and case-specific guidelines, technical solutions, and an empirical knowledge base that details the impact and harms of AI technologies to: “support multi-disciplinary bottom-up research and development in AI Ethics, particularly in commercial development contexts currently closed to external scrutiny.” (Mittelstadt, 2019, p.509)

Furthermore, the author proposes the formal recognition of AI development as a profession, particularly for developers of high-risk AI systems, akin to other high- risk professions. This would involve licensing practices, initially targeting developers working in public sector applications such as facial recognition for policing. Additionally, Mittelstadt suggests putting attention to developing organisational ethics, addressing the ethical challenges at the level of AI businesses and organisations, and not just individual developers: “The legitimacy of particular applications and their underlying business and organisational interests remain largely unquestioned. […] Developers will always be constrained by the institutions that employ them.” (Mittelstadt, 2019, p.510)

Overall, Mittelstadt criticised the technocentric idea that ethical challenges in AI can be addressed solely through technical fixes and “good design,” under the guidance of set ethical principles. This is because ethics is not a destination:

> “Ethics is not meant to be easy or formulaic. Intractable principled disagreements should be expected and welcomed, as they reflect both serious ethical consideration and diversity of thought. They do not represent failure, and do not need to be ‘solved’. Ethics is a process, not a destination.” (Mittelstadt, 2019, p.510)

He concludes that AI Ethics should be seen as a process involving continuous engagement with complex ethical debates, rather than seeking to simplify these debates into computable and implementable concepts. He calls for recognising ethics as an ongoing process, rather than a destination achievable with technical solutions or an adherence to ethical principles.

Although Mittelstadt does not explicitly call this a sociotechnical approach, some hints show why his argument supports a sociotechnical view of AI ethics. First, he recognises the importance of understanding AI ethics not only as a professional standard but also as an organisational change of paradigm. He also stresses the importance of integrating stakeholders and supporting bottom-up development.

Thus, as noted by Mittlestadt, many AI ethics initiatives have formulated guidelines that mirror the professional codes of ethics found in traditional professions, focusing on the conduct and principles that should guide individuals in their professional roles. This, I stress, can easily support technocentric solutions focused on how developers tackle a specific problem or how well they follow checklists, governance, and accountability demands.

Accordingly, most of these guidelines criticised by Mittlestadt, do not thoroughly scrutinise the legitimacy of specific AI applications or the business and organisational motivations behind them. This narrow focus tends to shift the discussion toward instances of misconduct by individuals, diverting attention from the broader ethical cultures, endemic within the organisations —i.e., disrupting a broader context for understanding the ethical aspects involved in AI development.

In practice, this disruption limits the narrative of AI systems as sociotechnical systems, that is, understanding AI merely as circumstantial tools that have a technical objective that needs supervision to "comply" with ethical expectations. In the case of bias, which is the focus of this work, this can translate into isolationist mitigation strategies that conform with the ethical principle of fairness defined through a bias-centric view of fairness, that is as "non-discrimination" and "bias prevention." But, as Mittlestadt argues, for AI ethics to be genuinely impactful, it must address the systemic issues present within the institutions that shape AI development.

Thus, it is necessary to recognise that individual developers operate within constraints set by their employers and their context. This calls for a broader approach to examining and addressing the collective ethical responsibilities of AI development, which requires caring not just for the individuals within them but encompassing the ethics of the entire AI ecosystem. This extended scope entails acknowledging that ethical practices must be embedded at all levels, from individual developers to the organisations that drive the AI industry, as well as other stakeholders, elements, factors, and incentives influencing AI development.

Mittlestadt also mentions the importance of seeing ethics as a process of continuous engagement, criticising the lack of inclusive design and (although not with this terminology) the prevailing technocentric solutions derived from principles in AI guidelines. For him, it is a misconception to believe that longstanding and complex moral issues can be adequately addressed with simplistic solutions. Ethics involves complex deliberations that should represent a deep engagement with a diversity of perspectives and experiences. Ethical debates are not indicative of failure, says

Mittlestadt, they need not be "resolved" because "ethics is a process, not a destination." (Mittlestadt, 2019, p. 510)

What I draw from Mittelstadt's criticisms and analysis is the following. Because we cannot guarantee ethical AI from principles alone, it is important to establish mechanisms that allow ethical practices to flourish. A particular challenge related to providing and supporting these mechanisms is both the sociotechnical intricacies of AI systems and the defiant absence of professional codes, institutionalised frameworks, and regulatory norms that provide a clear framework for AI developers. But this takes time, and aside from establishing general regulations, principles, and technical tools, solving issues about bias in AI could benefit from some of that ecosystemic robustness Mittlestadt alludes to.

Therefore, given that I agree with Mittelstadt's characterisation of ethics as a process rather than a destination, the Bias Network Approach proposal I will develop should include an open and iterative process that will allow developers to embrace the reflective process not as an academic exercise or practical standard for implementing an ethical framework, but as an internalised professional response to real and pressing bias-related challenges. Thus, it will serve as a call to action for AI developers to take a more integrated and holistic approach to their work, acknowledging the scope of their duties in connection to AI bias based on their role within the AI's sociotechnical ecosystem.

2.2.2 From principles to a situation-sensitive ethical approach.

Following a similar line of argument, Thilo Hagendorff argues that "AI ethics—or ethics in general—lacks mechanisms to reinforce its own normative claims." (Hagendorff, 2020, p. 99) One could immediately criticise this as a hasty claim, as having an enforcement mechanism is not what ethics should be doing nor what we should expect from it, that is a job for regulations and laws.

However, what seems to be the substantial point Hagendorff puts forward, is that the enforcement of ethical principles often results in institutions formulating their own ethical guidelines, creating the perception of internal self-regulation as a sufficient standard. In other words, ethical principles can easily be used as a cover for ethical compliance. Hagendorff highlights that the trend of adopting ethical principles, particularly in the private sector, fosters a reluctance to introduce more stringent regulations.

Essentially, the author suggests that the current state of affairs allows organisations to maintain the status quo by publicly adopting ethical stances without enacting substantial changes in their current practices, a phenomenon known as ethics washing.[6] This leads to a fundamental question posed by the author: "To what extent are ethical objectives genuinely implemented, as opposed to being mere expressions of good intentions?" (Hagendorff, 2020, p.100)

To analyse this issue, Hagendorff reviews 22 key ethical guidelines through a structured literature review.[7]

Based on this review, he criticises the lack of diversity and adaptability in AI ethics guidelines. He claims that guidelines cannot be universally applied across various contexts because they are too broad:

> "In general, ethical guidelines postulate very broad, overarching principles which are then supposed to be implemented in a widely diversified set of technical and economic practices, and in sometimes geographically dispersed groups of researchers and developers with different priorities, tasks and fragmental responsibilities." (Hagendorff, 2020, p. 111-2)

This broadness, says Hagendorff, means that "ethics […] operates at a maximum distance from the practices it actually seeks to govern." (p. 112)

[6] Recurrently used in the field of AI or technology ethics to refer to "[the] support of deregulation, self-regulation or hands-off governance, [where] "ethics" is increasingly identified with technology companies' self-regulatory efforts and with shallow appearances of ethical behavior" (Bietti, 2020, p.210).

[7] Hagendorff makes the following description of the methodology: literature review was conducted across various databases using AI ethics-related terms, examining the first 25 results per platform, and excluding duplicates and documents over five years old. Additional materials were identified through references in these sources. The process favoured English-language, Western publications and prioritized documents that provided a broad perspective on AI ethics, excluding national-specific reports except for globally influential ones.

As a result, the broad nature of ethical guidelines can lead actors, like developers, to avoid integrating robust ethical standards into their practices, resulting in the shifting of ethical responsibility onto others.

Hence, for Hagendorff, the minimal practical impact of AI ethics guidelines presents a significant challenge: closing the divide between high-level ethical principles and the specific technical actions required for their actual implementation.

He also notices that the guidelines reviewed use the term “AI” as a generic label for a wide-ranging set of technologies, without offering detailed explanations. “AI” covers a vast spectrum of applications, yet there is a noticeable lack of in-depth technical discussion linking the major ethical guidelines he examined.

To make a transition and fill in the gap between ethical theory and technical practice, Hagendorff suggests that ethics must evolve into “microethics,” adjusting its level of abstraction to engage meaningfully with technical disciplines and the practicalities of AI development:

> “On the way from ethics to “microethics”, a transformation from ethics to technology ethics, to machine ethics, to computer ethics, to information ethics, to data ethics has to take place. As long as ethicists refrain from doing so, they will remain visible in a general public, but not in professional communities.” (Hagendorff, 2020, p.111)

The author explains that a very direct way to improve guidelines would be to offer a supplement of technical explanations. This is intended to avoid deducing “concrete technological implementations from the very abstract ethical values and principles.” (Hagendorff, 2020, p.111) With this, Hagendorff means that a substantial change in the level of abstraction must happen —if AI ethics aims to have a certain impact and influence on AI development.

As an example of this transition into “microethics” Hagendorff references Gebru et al.’s (2018) paper. There, the authors introduce standardised datasheets with a list of properties for different training datasets. The point is that practitioners can check documenting practices and the composition of ML datasets. As declared by the authors of the datasheets, these documentation efforts are intended to enable dataset creators and consumers to engage with a more careful reflection about “the process of creating, distributing, and maintaining a dataset, including any underlying assumptions, potential risks or harms, and implications of use.” (Gebru et al., 2018, p.2) Thus, the datasheet provides a set of questions that can lead to a broader sociotechnical consideration. For example, some of the questions for the workflow, as presented by its authors, include:

- For what purpose was the dataset created?
- Was there a specific task in mind?
- Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?
- What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?
- Who was involved in the data collection process (e.g., students, crowdworkers, contractors)?

Hagendorff recognises, however, that:

> “[…] regardless of the fact that normative guidelines should be accompanied by in-depth technical instructions […] the question still arises how the precarious situation regarding the application and fulfilment of AI ethics guidelines can be improved.” (Hagendorff, 2020, p.112)

Accordingly, he suggests that a shift in AI ethics from a predominantly deontological approach, which relies on universal principles and rules, to one that incorporates virtue ethics, focusing on individual character and moral intuitions, is necessary.

Referencing Boddington (2017), Hagendorff emphasises that quite often ethical guidelines are perceived as “something whose purpose is to stop or prohibit activity, to hamper valuable research and economic endeavors (Boddington 2017)” (Hagendorff, 2020, p.112). Against this “negative” account of ethics, he suggests that the purpose of ethics is the exact opposite: “broadening the scope of action, uncovering blind spots, promoting autonomy and freedom, and fostering self-responsibility” (Hagendorff, 2020, p.112-3) —for which adopting a virtue ethics approach can help.

For the author, adopting virtue ethics means encouraging the cultivation of moral character trained to discern the wider implications of technological development whilst acting responsibly. In his argument, Hagendorff calls for institutional changes that support these ethical shifts, such as legal frameworks for technology auditing, official institutions to attend AI-caused harms, and expanded university curricula that include ethics in technology. The overall aim is to not restrict innovation but to enable more ethically aware and responsible action within the field of AI.

Hence, Hagendorff ́s central arguments support that AI ethics must go beyond checkbox-style guidelines and embrace a more nuanced, virtue-based approach that prioritises situational sensitivity, personal dispositions, and responsible autonomy. This perspective does not force a one-size-fits-all application of principles but instead appreciates the unique circumstances of each case and the technical specifics involved.

For Hagendorff, AI ethics should empower moral agents with the knowledge and empathy needed to make responsible decisions, rather than strictly enforcing adherence to normative principles. To achieve said goal, says Hagendorff: "AI ethics should not try to discipline moral actors […] but emancipate them from potential inabilities to act self-responsibly on the basis of comprehensive knowledge." (Hagendorff, 2020, p.114). Thus, Hagendorff's proposal raises various concerns, including how we may realistically migrate to microethics, or whether the diagnosis of AI ethics being primarily based on deontological approaches is as ubiquitous as he claims. Here, however, I will focus on some of the problems that Hagendorff emphasises regarding AI ethics in general, to characterise elements that my BNA proposal should address as a sociotechnical intervention.

First, Hagendorff emphasises the importance of self-responsibility, calling ethics to empower moral agents to make responsible justified decisions. As I will argue in Chapter 6, the adoption of the BNA will contribute to the development of self- and active responsibility among developers, by promoting a reflective and context- sensitive viewpoint. Thus, by adopting a more nuanced and context-sensitive approach like the BNA, developers can evolve a stronger sense of self-responsibility, thereby counteracting the tendency to shift ethical responsibility onto others, as discussed by Hagendorff.

Another critical point raised by Hagendorff is the need to change the level of abstraction of AI ethics because ethics should have a certain impact and influence on AI development. Part of his proposal to achieve a less abstract solution is to adopt a virtue ethics approach, which translates into encouraging the cultivation of moral character trained to discern the wider implications of technological development whilst acting responsibly. In his argument, Hagendorff also calls for institutional changes that support these ethical shifts, such as legal frameworks for auditing, and even expanded university curricula that include ethics in technology.

Although I agree with the overall objective of implementing these types of changes, I will focus particularly on one: the cultivation of moral character for AI developers. With my BNA proposal, I wish to offer AI developers —who are willing to exercise their virtues— an approach that allows them to make prudent choices when evaluating biases, i.e., to be able to discern and make decisions embracing the complexities of the AI ecosystem.

As regulatory frameworks and other "highly abstract" guidance such as ethical protocols or AI principles keep emerging, developers should be able to grasp the essential distinctions provided by these normative guidelines to develop contextual awareness, and collaborative ethical evaluations that can help them achieve practical wisdom. Accordingly, the BNA will support this, even if results in achieving this goal are rudimentary at first.

2.2.3 The uselessness of AI principles.

More recently in this continuous criticism of the practical limitations of AI ethics guidelines, Luke Munn argues about the uselessness of AI ethics for providing "meaningless principles, isolated principles, and toothless principles […] a gap between principles and practice" (Munn, 2023, p.869-870). By meaningless principles, he refers to the proliferation of AI ethics guidelines that offer abstract and ambiguous guidance, lacking specificity, and leaving room for existing practices to endure i.e., maintain the status quo in an industry that he identifies as systematically ignorant of ethics.

Ethical AI principles often reference the vices of tech culture, he says. Referencing Rességuier and Rodrigues (2020), Munn emphasises that AI ethics is “toothless” because ethics is understood as a replacement for regulation, which results in ethical frameworks that “set normative ideals but lack the mechanisms to enforce compliance with these values and principles.” (Munn, 2023, p.871) The result of this misplaced expectation is a gap between high-minded principles and technological practice.

Both authors previously mentioned, Mittelstadt and Hagendorff, shared this very same critique, pointing out vague guidance, a lack of connection between principles and practices, and the limited enforcement provided by guidelines —which shows how these criticisms continue to be relevant through the years.

However, in Munn’s argument, he highlights the reluctance of engineers to engage with ethical questions, which is symptomatic of a larger, more pervasive problem within the tech industry, which is making “unethical AI […] the logical byproduct of an unethical industry.” (Munn, 2023, p.871) Munn points out a fundamental cultural and ethical gap within the industry, which results in AI systems that mirror these deficiencies. He illustrates this by referencing prevalent attitudes in AI development where biased AI models are often dismissed because they just need tweaking: “To suggest that an AI model is “biased” and only needs to be tweaked is to adopt a far too narrow scope, missing out on broader or more systemic issues.” (Munn, 2023, p.871).

Furthermore, Munn argues that even when this gap between principles and practices is acknowledged, and principles are “operationalised,” the translation from complex social concepts to technical rulesets is non-trivial. Therefore, blind (to broader context elements) or rushed efforts to overcome this can be counterproductive. For example, he notices that:

> “[…] researchers and companies have aimed to make ethical values feasible and actionable in real-world settings […] However, operationalizing AI ethics promises to be difficult or even impossible, a daunting challenge underestimated by a technically focused industry and even by ethicists.” (Munn, 2023, p.872-3)

The problem Munn points at is the outcomes of this trivial understanding and implementation of AI ethics in industry and research, resulting in a simplification of AI ethics —an overall lack of robustness that makes it "useless." Munn explains that there is a casual attitude to dismiss problems about crucial and complex issues like fairness and privacy as if they were easily resolved:

> "These are highly contested issues, with high stakes. What is fair and who gets to decide it? […] And how might fairness play out differently in different contexts and conditions? These are complex questions which have shifted substantially over time and which intersect with race, gender, and culture." (Munn, 2023, p.873)

Munn offers two avenues to address this, a broader and a narrower approach. He introduces the notion of "AI justice," supporting broader and holistic scrutiny of ethical matters in AI. Munn argues that AI justice recognises AI systems as not morally neutral but as entities shaped by the societal contexts in which they function. This perspective embraces the intersectionality of various social and political factors, including historical disparities, race, gender, and cultural dynamics —supporting sociotechnical approaches to AI, recognising broader context and external elements influencing and being influenced by AI systems.

Thus, in practice, Munn suggests that to implement AI justice, we should engage with marginalised communities that are most affected by AI, promoting their well-being, and avoiding the exacerbation of historical inequities. In a way, it challenges the status quo that ethical guidelines allow to maintain, problematising principles such as fairness, which he argues has been taken for granted:

> "Historically fairness has been defined by hegemonic groups in ways that perpetuate their advantage: far from being "common sense," fairness is always historical and cultural with major racialized and gendered dimensions […] At a concrete level, it may mean organizations engaging with groups that bear the brunt of AI impacts but are not typically consulted: children, people of color, LGBTQIA+communities, migrants, and other groups." (Munn, 2023, p.873-4)

On the other hand, Munn discusses a narrower approach that focuses on specific issues within AI, such as accuracy, alignment, and impacts. This approach involves concrete actions like creating balanced datasets to address bias and improving the transparency of the model. This narrow view emphasises the material and measurable aspects of AI ethics, moving away from vague or more conceptual notions of ethics, and instead focusing on accountability and documentation procedures. For this, Munn says that oversight and auditing are key, considering that several data governance, AI production, and design tools can help developers to:

> "[…] audit their work at each stage and see how well it matches organisational principles […] [providing]better oversight about the kinds of decisions that are being made and the kinds of (potentially harmful) consequences that may result." (Munn, 2023, p.874)

For Munn, addressing the ethical challenges of AI requires a more nuanced and multifaceted approach that goes beyond ethical principles. It calls for transparency, accountability, oversight, and concrete actions to reshape AI in a way that aligns with ethical values and avoids harmful consequences. Ultimately, it highlights the importance of involving various stakeholders, including developers, governments, and professional societies, without relying on ethical guidelines as the main resource to enforce or implement AI ethics.

In line with the criticisms of authors before him, Munn refers to the lack of structural integrations of AI ethics into the development context. Munn criticises the proliferation of AI ethics guidelines for being too abstract and failing to provide specific guidance, which leaves room for existing unethical practices to persist. He suggests that the tech industry is systematically ignorant of ethics, leading to the development of AI systems that do not adequately incorporate ethical considerations. In part, this last issue occurs because of another systemic problem, which is the lack of appropriate ethical training.

And, as I have mentioned before, the BNA will be presented as a sociotechnical approach to this lack of ethical training, seeking to support these gaps through interdisciplinary collaboration.

In this sense, I believe approaches like the BNA proposal I will develop can contribute —to some extent— to overcoming the criticised "uselessness" or "toothlessness" of AI ethics. The BNA will be designed to translate part of the theoretical discussion, in this case about bias, into specific practices that help AI developers deal with ethical queries.

This could, as well, if adopted institutionally, combat the institutional cultural and ethical gaps identified by Munn, particularly within the tech industry, where there is a reluctance to deeply engage with ethical questions.

My proposal will encourage the opposite, by establishing an interdisciplinary dialogue to actively engage AI developers with broader ethical inquiries about bias.

2.2.4. Conclusions for section 2.2: bridging the gap between abstract ethical principles and practice

The critiques of AI ethics principles offered by Mittelstadt (2019), Hagendorff (2020), and Munn (2023) converge on the limitations of current ethical frameworks in adequately addressing the complexities of AI development. These scholars point to the insufficiency of abstract principles to resolve the nuanced ethical issues that arise within AI systems, while also critiquing the performative nature of existing guidelines.

Their work collectively highlights the need for a paradigm shift towards a more situated, systemic, and reflective approach to ethics in AI—a shift that the Bias Network Approach (BNA) operationalises through a sociotechnical intervention.

In what follows, I highlight the similarities between my proposal and the critiques raised by these authors, showing how my work responds to their core concerns. I also outline key differences, particularly focusing on how my approach addresses the gap they identify between abstract ethical principles and their practical implementation. Specifically, I argue that my proposal offers a tangible intervention that effectively bridges this divide, moving beyond the theoretical critiques to operationalise ethical reflection and decision-making in practice.

Notwithstanding, before examining this, it is worth noticing that although my proposal will contribute to reduce this practical gap, it is not necessarily a solution to a specific issue criticised by these authors, particularly Hagendorff, when he calls AI ethics principles “lacking reinforcement”. As it will be discussed in further chapters, the BNA will not enforce the adoption of a practical and contextual ethical approach, despite offering one. For example, if a company wants to be unethical - e.g. if they put profit before public safety, then the BNA may not help. They may just ignore it. However, this bridge between abstract principles and practice can be provided by the BNA to cases where developer teams or institutions want to behave ethically. Having this in mind, let us now turn to aspects of the BNA that do contribute to reducing the gap.

Mittelstadt (2019) stresses the importance of viewing ethics as an ongoing process of continuous engagement, rather than a destination defined by adherence to fixed principles. He critiques technocentric solutions for their inability to grapple with the enduring and complex nature of moral challenges in AI. Ethics, for Mittelstadt, demands a deep engagement with diverse perspectives and cannot be reduced to simplistic resolutions. His emphasis on ethics as a process aligns with the BNA’s design, which integrates iterative and context-sensitive reflection into AI development practices.

The BNA moves beyond ethical checklists by encouraging developers to engage with bias-related challenges not as static problems, but as evolving issues tied to the sociotechnical ecosystems in which they operate. This dynamic and reflective methodology not only complements Mittelstadt’s critique but also empowers developers to adopt a proactive, integrated, and holistic approach to addressing bias as an embedded challenge within AI systems.

Hagendorff (2020) critiques the negative perception of ethics as merely a restrictive force that stifles innovation and progress. Instead, he argues that ethics should empower developers by uncovering blind spots, promoting autonomy, and fostering self-responsibility. His call for a virtue ethics approach, which emphasises the cultivation of moral character and situational sensitivity, resonates deeply with the objectives of the BNA.

The BNA integrates this call by offering developers structured opportunities for reflection, documentation, and interdisciplinary collaboration, thereby ensuring that accountability becomes a cornerstone of ethical practice. Thus, the BNA addresses both individual responsibility and systemic challenges about bias in AI ethics.

Finally, Munn (2023) critiques the tech industry's reluctance to deeply engage with ethical questions, attributing this to systemic cultural and institutional gaps. He calls for a nuanced and multifaceted approach to ethics that prioritises transparency, accountability, and oversight while moving beyond the abstract and vague recommendations often found in ethical guidelines. Munn's emphasis on involving diverse stakeholders, including developers, governments, and professional societies, aligns with the BNA's commitment to interdisciplinarity and sociotechnical awareness.

The BNA addresses this critique by explicitly situating bias within its broader societal, cultural, and political contexts. Drawing on concepts from feminist epistemology, such as situated knowledges (Haraway, 1988) and strong objectivity (Harding, 1992), the BNA encourages developers to reflect on their positionality and the systemic inequalities embedded in AI systems. This practical focus ensures that ethical reflection is not treated as a separate or optional component of AI development but as a core part of the decision-making process.

Collectively, these critiques emphasise that ethical challenges in AI are too complex to be addressed by abstract principles alone. As Mittelstadt, Hagendorff, existing guidelines fail to provide the actionable and systemic guidance necessary for meaningful ethical engagement. The Bias Network Approach offers an alternative to overcome these limitations by operationalising a sociotechnical orientation that integrates interdisciplinary collaboration, iterative reflection, and context-sensitive analysis into AI development practices.

By integrating these perspectives, the BNA advances beyond the narrow confines of "algorithmic fairness" or "bias-centric mitigation strategies," incorporating considerations of societal power dynamics, cultural influences, and interdisciplinary insights. This positions the BNA as a practical and transformative framework that not only addresses the critiques of existing ethical principles but also provides a robust foundation for fostering genuine ethical practices in AI development. In doing so, the BNA represents a significant step towards realising the paradigm shift that Mittelstadt, Hagendorff, and Munn collectively call for, ensuring that ethics becomes an integral and impactful component of AI development.

2.3 Another way to conceptualise societal bias.

Now, I will turn to analyse a specific proposal to reconceptualise the concept of societal bias, and that embraces a sociotechnical view. Mike Zajko (2021) argues that prevalent approaches to addressing bias in AI can be categorised as "conservative," meaning they maintain the existing societal order instead of challenging and disrupting systemic forms of inequality.

In contrast, Zajko argues in favour of a "radical" approach that incorporates social theory and embraces a broader societal context to effectively engage with bias-related concerns.

To set the discussion, Zajko starts by recognising that there is an increasing expectation for AI practitioners to consider the ethical and political ramifications of their work, entering discussions traditionally occupied by social sciences and humanities. Debates about societal bias, have primarily prompted questions about discrimination and unfairness, given that AI systems often embed the prejudices of the societies from which their data is gathered. However, there is a deficiency in tackling these interdisciplinary challenges related to bias, as Zajko notices:

> "Technologists are often poorly prepared for these considerations, and dominant paradigms in data science have been criticized as narrow technical approaches to social problems, necessitating involvement from additional perspectives." (Zajko, 2021, p.1047)

Among the multiple factors that create this status quo for AI development (as also pointed out in section 2.2), Zajko highlights the prevailing logic of capitalism, institutionalised cultures within academia promoting formalism (e.g., technocentric practices that, for example, use formalisations of fairness to solve issues about bias, (what I called bias-centric view of fairness), the alignment of projects to the interests and motivations of the sources of funding (pervasive incentives), as well as general issues with lack of diverse working environments in private and public sectors. Current approaches to addressing bias in AI criticises Zajko, tend to rely on data- driven technical solutions failing to consider the social context in which AI is developed and unfolds.

Therefore, to challenge this status quo Zajko proposes to embrace interdisciplinarity, as this can extend our understanding of bias within AI facilitating an articulation of societal values. The author claims that given that politics is inherently about power, traditional computing and data science approaches tend to be politically conservative, reinforcing existing power structures.

Accordingly, to Zajko, the overreliance on conservative approaches leads to vague notions of "doing good" which do not encompass objectives related to social justice or other explicit political goals. In this regard, conventional strategies to address bias and formal mitigation approaches are insufficient for achieving these objectives. To approach this issue, Zajko notices that:

> "AI has the potential to disrupt various institutions and social processes but is typically used as a tool to reinforce the status quo and benefit those at the center, rather than the margins." (Zajko, 2021, p.1048)

To develop his radical approach, Zajko makes a critical examination of societal bias, interpreting it as an issue of social inequality. The author argues that conventional definitions of bias often concentrate on the precision of AI predictions or classifications, e.g. if they discriminate, the evaluation of the bias is subsumed to the performance of the system. However, says Zajko, "societal bias indicates some undesirable state of affairs" (Zajko, 2021, p.1048). Accordingly, for Zajko, to talk about societal biases in AI, it would be better to study types of social inequality, as this can yield more effective ways of conceptualising these issues, enabling us to "imagine futures that are not limited to the removal of bias." (ibid) This perspective entails situating AI systems within the context of existing social infrastructures and questioning how these systems may perpetuate, modify, or renovate these structures.

Zajko's radical approach, therefore, requires us to consider social theories and frameworks to comprehend and tackle the underlying inequalities intertwined with bias in AI systems. By adopting this broader perspective, the goal is to move beyond merely addressing the symptoms of bias and instead confront the systemic roots of inequality. This theoretical integration enables, according to Zajko, a more comprehensive understanding of the complex dynamics at play and provides a foundation for formulating alternative approaches that can genuinely challenge and transform the existing social order.

In summary, Zajko argues for the adoption of a radical approach that embraces social theory and a broader societal context to address bias in AI. By challenging the conservative tendencies that reproduce existing inequalities, this approach seeks to foster a transformative and equitable framework to confront the conservative status quo. To enact this transformation, Zajko argues that:

> "[…] this interdisciplinary engagement needs to happen early in the development of AI systems; it is not simply a matter of adding the missing social context to an already-formulated problem. When we begin by naming and analyzing the social structures we find problematic, we can think about ways of changing them or addressing their harms" (Zajko, 2021, p.1050).

Hence, to Zajko, social science research offers the vocabulary and framework needed to identify and address problems about AI bias. For an interdisciplinary approach to have a tangible impact, however, it is crucial to incorporate it from the very beginning of AI system development, rather than attempting to adapt social circumstances to pre-defined problems after the fact.

Delaying the incorporation of these factors, says Zajko, increases the likelihood that the socio-technical system may contribute to the problem is trying to solve, making subsequent remedies superficial. Then, to fully exploit the benefits of an interdisciplinary approach, Zajko argues that we should begin by considering the fundamental question of what the optimal result would be, considering our comprehension of the factors that contribute to inequality. He suggests that:

> "The most obvious way that we can move beyond the negative orientation of 'removing bias' is to specify social inequality as the problem, and equality or equity as a desirable outcome to work towards." (Zajko, 2021, p.1051)

He argues that the language of societal bias currently used in AI does not help us examine how inherently unjust societal structures like capitalism, patriarchy, or colonialism function and impact the field.

Zajko references Hoffman to stress this point. For Hoffman, the notion of biases does not focus on "the normative conditions that produce—and promote the qualities or interests of—advantaged subjects." (Hoffmann, 2019, p. 907) Following this, Zajko offers what he calls an "interdisciplinary contribution," claiming that the language of societal bias in AI could "benefit from being transformed or replaced by more elaborated concepts in social theory related to inequality." (Zajko, 2021, p.1050)

Zajko states that the way to move beyond negative and simplistic orientations of removing bias is to make social inequality the core concept:

> "[…] is better understood as the intersection of different structures of inequality, as named and analysed by scholars in the social sciences and humanities prior the current era of machine learning." (Zajko, 2021, p. 1054)

Consequently, Zajko's proposal incorporates sociological concepts such as intersectionality, structural inequality, and critical race theory, as a conceptual change to understand societal bias. By integrating these elements into a broader contextual framework, Zajko claims that it is possible to enhance efforts aimed at mitigating systemic inequalities and promoting social justice. For the author, the incorporation of social theory enables a comprehensive understanding of the societal, cultural, and political factors that shape the development and deployment of AI technologies. Thus, the social theory critical lens, makes it possible to recognise and analyse how AI systems can (sometimes inadvertently) perpetuate and amplify existing social inequalities.

Now, regarding how this proposal should be translated into practice, not much is said. For example, in relation to the integration of that "intersection" of social structures, Zajko says that AI researchers and developers:

> "[…] need to be able to supplant terms like 'racial bias', which restricts further analysis, with theories of racial inequality that open up further avenues for analysis —including how race intersects with other social hierarchies." (Zajko, 2021, p. 1053)

Doing so, says the author, opens the possibility of specifying goals beyond accuracy, efficiency, or bias reduction. This intersectionality, however, is not defined or examined further as a concept, even though Zajko recognises its richness and tradition within social sciences and humanities. Overall, he concludes that:

> "The concept of bias is limiting and should often be jettisoned, where more specific conceptualisations of inequality are available. Rather than being concerned over how socio-technical systems reproduce pre-existing biases, we can actually name what we want to avoid reproducing: identifying processes, structures, hierarchies and concepts." (Zajko, 2021, p.1054)

This call to stop using bias and instead talk about more specific instances of inequality is presented by Zajko as a way to push against conservative ideologies pervading the development of AI systems. What the author points out is that taking care of societal biases does not really solve the systemic societal problem we face and does not necessarily guarantee profound change.

2.3.1 Criticisms and benefits of Zajko's radical proposal.

Zajko proposes a change in how societal bias is understood in AI. But, as he presents it, this change would also require developers to understand profound social concepts:

> "Wherever racial bias […] is an issue, the least that a developer can do is to understand what race is and how racial inequality is structured in society. While this might seem like an obvious point, there is still an enormous amount of work being done in computing and data science to classify races, genders, emotional states, or potential for criminality, with only the shallowest ontological engagement with these phenomena." (Zajko, 2021, p. 1053)

The level of conceptual understanding this approach calls for may be too demanding. While it is reasonable to expect a baseline of awareness and knowledge from AI developers, to meet these expectations would necessitate significant changes in their educational and training processes, including deeper engagement with social, economic, and political (SEP) concepts —pointing out the profound cultural and institutional changes discussed in the previous section 2.2.

Imposing such a level of conceptual understanding, as suggested by Zajko, may impede the establishment of responsibility and accountability in developers. It is not necessarily evident that an in- depth grasp of concepts such as race, racial inequality, colonialism, ageism, gender inequality, etc., will enable developers to integrate this knowledge into their decision- making processes. Nor is it certain that such a profound level of knowledge is a practical or fair requirement to expect from them [developers].

Moreover, Zajko criticises the existing definition of societal bias saying that: "societal bias indicates some undesirable state of affairs, but without a basis for imagining what is desirable." (Zajko, 2021, p.1048) However, I disagree with this claim. It is not evident why a definition of societal bias should establish what is desirable.

As I will discuss in Chapter 3, this seems to fall outside the scope of what bias is. Zajko argues that to design improved AI systems, the process of imagining these "desirable futures" should prioritise the consideration of inequalities, instead of simply reacting to instances where biased data or decisions are identified. But this emphasis on inequalities, I argue, should not replace the acknowledgement of bias and its technical manifestations. Hence, I will argue that bias should not be jettisoned as Zajko claims but rather used to establish a comprehensive normative foundation that informs AI development.

Accordingly, I suggest that the existing categories of bias in AI (technical, cognitive, and societal, as I will describe in section 3.1) offer valuable insights into the perpetuation of pre-existing biases, including that of societal bias. They serve as a preliminary means to identify and categorise the types of biases that are being reproduced. Through this identification process, contextual information such as the origins of bias, can be obtained, thereby facilitating the task suggested by Zajko — that is, naming and monitoring the inequalities that require avoidance.

Now, to complement this initial identification, a more comprehensive understanding of bias and its underlying mechanisms can be achieved by integrating concepts of specific inequalities (the intersectional approach Zajko proposes). This second-order identification could go beyond simply recognising the manifestation of biases and delve deeper into understanding the intricate processes, structures, and hierarchies that enable bias to persist.

Such a nuanced approach is essential for developing effective interventions and addressing the root causes of bias in AI systems. However, it appears that Zajko's proposal may overlook the significance of the technical and systematic aspects that have been developed to address AI bias.

By emphasising the social dimensions of bias- related issues, there is a potential risk of neglecting the technical complexities and challenges to mitigate bias. Hence, a more comprehensive approach would consider the interplay between social, technical, and cognitive factors, recognising that bias reduction requires interdisciplinary collaboration and the integration of multiple perspectives as well.

Thus, sociotechnical approaches do require further context for analysing issues about biases. But, as I will propose, this should also include the technical aspects that ground AI development in practice.

Previously, I criticised the technocentrism, bias-centric views of fairness, and isolationism affecting how AI bias is being addressed. The primary shortcoming of technocentric and bias-centric viewpoints is their potential to overlook the wider context and the interplay of biases and inequalities. Therefore, it is important to acknowledge that while Zajko's comprehensive approach is conceptually robust, it might present practical difficulties for AI developers, especially those without a background in ethics or social sciences. Implementing such a demanding approach requires engagement with intricate conceptual frameworks and interdisciplinary knowledge.

Consequently, it is important to find an intermediate solution, one that does integrate part of the conceptual complexity and robustness suggested by Zajko, but that can be translated into AI developers' practice.

From this critical analysis of Zajko's proposal, I reflect on an essential aspect to consider in the development of AI systems: achieving interdisciplinary engagement, which entails collaboration between technical experts and scholars from diverse fields. This collaborative effort is crucial for integrating ethical considerations, social implications, and systemic factors into AI system design, thereby fostering ethical development.

The adoption of interdisciplinary collaboration, as presented by Zajko, requires a shift in focus, away from exclusively addressing bias mitigation, and into the problematic societal structures grounding societal bias.

In a more realistic context, however, since it may not always be feasible to completely alter these structures, understanding and acknowledging the influence and role biases in AI development becomes imperative —as well as making transparent these findings and communicating them to the AI community. To achieve this, interdisciplinary approaches allow for a more comprehensive examination of the broader sociotechnical context within which AI systems operate.

I would also like to clarify that my reservations towards highly conceptual and abstract approaches like Zajko's stem from the concern that in seeking interdisciplinary collaboration, it may be unreasonable to expect AI developers to shoulder the primary responsibility for dissecting the complexities of these ethical issues related to societal bias. Zajko himself concedes that addressing issues like overrepresentation in the criminal justice system requires tackling broader injustices and processes of criminalisation, which extend beyond simply devising a "fairer algorithm" and involve a deeper understanding of racial inequalities.

Thus, the kind of engagement with various forms of inequality that Zajko encourages should not necessitate discarding current, workable concepts of bias such as societal bias or racial bias —as he suggests, nor should it mandate that AI developers immerse themselves in the exhaustive study or recognition of theories related to these inequalities.

A more balanced approach is preferable, I argue, one that encourages cooperation among experts across various fields to identify societal and technical factors that contribute to biases within AI systems, promoting active interdisciplinarity collaboration.

I will suggest, in Chapter 4, that the BNA can provide a robust foundation to acquire sociotechnical insights about bias, as suggested by Zajko, but in a more accessible way for AI developers to engage with interdisciplinary intersectionality.

Sociotechnical advancements in AI ethics might benefit from a less radical approach, focusing on a collaborative effort involving experts from multiple disciplines to comprehensively explore the web of societal structures, but without dismissing technical components influencing AI systems' biases.

2.4 A sociotechnical systemic approach to bias.

So far, I have introduced criticisms of AI ethics that emphasise sociotechnical changes to bridge gaps in how ethical principles are influencing AI development. I have also discussed a proposal to conceptually change the term "societal bias," criticising it but also gathering insights about the necessary elements that need further development to properly address bias in AI. From all these views, I highlighted something worth considering for the design of the Bias Network Approach (BNA).

In this final section, I will examine a proposal of a sociotechnical systemic approach to bias by Draude et al. (2019), which will serve as a direct conceptual foundation for the BNA, in addition to the insights previously highlighted.

Draude et al. (2019) propose a sociotechnical systemic approach to bias arguing that technological development is intrinsically linked with power dynamics and inequalities —aligned with the views provided by the other sociotechnical approaches discussed here. However, despite recognising that these interactions of societal dimensions and technology are a longstanding discussion, the authors emphasise there is a pressing need for translational work to implement sociotechnical approaches, which recognise the co-construction of society and technology.

This perspective on translational work is critical for understanding how biases are not only replicated but also how they can be systematically addressed through interdisciplinary collaboration —a goal I have already identified as part of the basis for developing the BNA. In this respect, Draude et al. (2019) argue that:

> "Technological development is closely related to power dynamics and social, economic and political inequalities. […] Computer scientists and engineers oftentimes are not educated accordingly. Also, the complexity of the social world provides a challenge for any development process. Sociotechnical approaches (Bijker & Law, 1992) offer solutions by postulating the co- construction of society and technology. They require translational work that must be done between and across disciplines (cf. Lin, 2012)." (Draude et al., 2019, p.326)

Grounded on gender and diversity studies as critical tools for analysing and mitigating algorithmic bias, the authors pose key questions about the nature of bias. Their proposal for addressing AI bias incorporates concepts from feminist philosophy of science, that fit organically within sociotechnical system theories, to offer what they call a sociotechnical systemic approach to biases.

Their proposal emphasises the need for translational work across disciplines and offers specific suggestions for researchers and practitioners on how to account for social inequalities in the design of algorithmic systems, urging for a systemic approach to bias that engages with the complexities of the AI ecosystem.

To contextualise their discussion, the authors explain that their sociotechnical approach to AI recognises biases not as entities opposed to humans but as present within human experiences. More specifically, they notice two distinct levels of interconnection.

The first level involves social inequalities perpetuated through automated decisions. The second level involves the cultural role of algorithms in organising and prioritising information and activities, contributing to what Striphas calls "algorithmic culture." (Striphas, 2015) This culture encapsulates human thought and human activity within the realms of computation. Consequently, algorithms emerge as central, formative structures, instrumental in shaping human existence.

For example, Draude et al., mention that search engines can become political because the algorithms they use to rank and display websites inherently make choices about which sites become prominent and which remain unseen. These engines can systematically prioritise results, inherently biasing the user's experience. This outcome is a product of the purpose behind the search engine and the decision-making criteria embedded in its algorithms, which are shaped by the practices and contexts of development.

Over the years this discussion about biases has evolved, but some of the initial ethical challenges regarding bias in AI systems remain. As I have stressed in previous sections of this Chapter and Chapter 1, a prevalent approach to bias (and to AI ethics more broadly) has been technocentric.

Although sometimes grounded on ethical motivations (i.e., explaining unjust results of algorithms and offering fixes to prevent them), it is nowadays recognised that many of these initial solutions fall short of confronting and fixing profound inequality issues.

For example, we now know that the effectiveness of fairness criteria like demographic parity and equal opportunity in algorithms becomes futile if protected attributes are redundantly encoded through proxy variables. This suggests a significant limitation in the way fairness is algorithmically operationalised, as it may not account for more subtle forms of bias embedded within the data, hence creating a recurrent challenge developers must be aware of and confront.

Other initiatives providing open- source tools like Aequitas and AI Fairness 360, represent significant advances towards the democratisation of algorithmic accountability, empowering users to identify and report biases, potentially leading to more inclusive and equitable AI systems. But, once again, their effectiveness is contingent on the users' ability to understand and engage with complex statistical criteria and technical and societal concepts of fairness, which may not be universally accessible.

Hence, Draude et al.'s proposal starts from the criticism of a significant reliance on technical tools for accountability, as emphasised by Kroll et al. (2017), for not fully addressing the intricate sociotechnical dynamics that give rise to biases in AI.

This, again, highlights the need for a multidisciplinary approach that incorporates broader perspectives, suggesting that technical interventions alone are insufficient. Addressing AI bias, therefore, requires a holistic approach that considers the broader societal and cultural structures that shape AI development, which is what Draude et al. suggest can be achieved by reframing the question of bias through an intersectional gender studies perspective:

> "Inequalities are indexed on social categories (race, class, gender, ethnicity, sexuality and dis/ability) that can work simultaneously and form complex patterns of power and hierarchy– Crenshaw (1989) called this "intersectionality." Feminist epistemology shows that these patterns in turn influence what counts as valid knowledge and as objective fact, whereas the binary, hierarchized view of gender relations has historically served as a ready-made model for interpretation of scientific findings." (Draude et al., 2019, p. 330)

Based on the concepts of situated knowledges (Haraway, 1988) and strong objectivity (Harding, 1992, 2015), Draude et al. propose a way to rethink bias from that intersectionality.

Donna Haraway developed the concept of "situated knowledges" to challenge the idea of objectivity in scientific methods. She argues that all knowledge is "situated" and constrained by the unique circumstances, backgrounds, and context of its creators, alluding to a positional standing for scientific knowledge:

> "Situated knowledges are about communities, not about isolated individuals. The only way to find a larger vision is to be somewhere in particular. The science question in feminism is about objectivity as positioned rationality. Its images are not the products of escape and transcendence of limits (the view from above) but the joining of partial views and halting voices into a collective subject position that promises a vision of the means of ongoing finite embodiment, of living within limits and contradictions of views from somewhere." (Haraway, 1988, p.590)

Thus, Draude et al. stress the fact that knowledge production —in this case through AI development, is unavoidably embedded in the positionality of AI developers and their context.

Furthermore, just like for Haraway, there are no neutral observers in scientific production, there should be no neutral AI developers. Draude et al., clearly stress the importance of situated knowledges specifically within sociotechnical approaches:

> "Our forms of vision allow us to see certain things and at the same time obscure others, while where we see from implicates us in the web of power relations and influences both how and what we can see. This partial perspective, if not reflected upon, becomes a problem in sociotechnical systems design whenever the developers take their perspective as representative of the end user. Thus, we need to re-think scientific knowledge production not as universal but rather as valid from a specific perspective or position that operates always within certain figurations of time, space and artefacts, which is to say: as situated knowledges." (Draude et al., 2019, p. 331)

Hence, the authors argue that the perspective of situated knowledges offers a strategic and systematic way to consider power disparities and varied viewpoints within a sociotechnical framework. Rather than claiming an unrealistic neutrality, their approach is presented as an enabler of a reflective understanding of knowledge.

Here is where they draw from Harding's standpoint theory and the notion of "strong objectivity." (Harding, 1992, 2015) Harding recurrently emphasises the "feminist attempts to transform the notion of objectivity so that it could function more effectively," (Harding, 2015, p.31) referring to the work of feminist philosophers and scientists like Karen Barad (2007), Heather Douglas (2009), and Helen Longino (1993).

Harding, however, argues that her standpoint theory proposal is slightly different because it starts from understanding knowledge production origins in the real world, where most scientific research is influenced by corporate interests and technically driven cultures. Accordingly, for Harding, the notion of strong objectivity requires that:

> "[…] the subject of knowledge be placed on the same critical, causal plane as the objects of knowledge. Thus, strong objectivity requires what we can think of as "strong reflexivity." This is because culturewide (or nearly culturewide) beliefs function as evidence at every stage in scientific inquiry: in the selection of problems, the formation of hypotheses, the design of research (including the organization of

> research communities), the collection of data, the interpretation and sorting of data, decisions about when to stop research, the way results of research are reported, and so on." (Harding, 1992, p. 69)

Harding establishes a standpoint where knowledge is socially situated, particularly including perspectives of marginalised or oppressed groups that have been historically apart from scientific knowledge —as they can provide more complete and less distorted insights into social reality. This is because the subject of knowledge must integrate both the perspective of the "outsiders," and the perspective of those in power to navigate society effectively.

Thus, the concept of "strong objectivity" criticises the traditional understanding of objectivity, which means assuming that the production of knowledge should come from a "neutral stance" i.e., a "weak objectivity."

Weak objectivity refers to the standard practices in scientific research based on neutrality and detachment, establishing that researchers can detach from their biases, perspectives, and social positions to produce neutral, unbiased knowledge.

To some extent, this criticism of weak objectivity relates to the technocentric views criticised in Chapter 1. The neutral objectives of scientific practice can promote a status quo where dominant groups' perspectives are adopted as the default or neutral standpoint, masking the biases and assumptions affecting them. For instance, if a field is dominated by a particular racial, gender, or cultural group, their perspectives and interests might be wrongly assumed to be universally objective or the default to understand a particular problem.

The same can apply to other standpoints influencing how biases are conceived. Thus, "weak objectivity" overlooks how societal values and power structures shape research questions, methods, and interpretations.

In other words, traditional notions of objectivity often lack a mechanism for critically examining how a researcher's positionality affects their work. This lack of self-awareness means that biases and assumptions can go unchecked.

Overall, Harding’s concept of strong objectivity calls for an acknowledgement of the researcher’s perspective and a systematic examination of how that perspective influences the research process. The goal of strong objectivity is to produce more accurate and comprehensive knowledge by recognising and accounting for the social and power dynamics that influence research, for which interdisciplinary interactions are key, as Harding explains:

> “The strong objectivity program argues that starting research from “outside” a discipline can enable the detection of the dominant values, interests, and assumptions that may or may not be widely prevalent, but which tend to serve primarily the most powerful social groups. “Dominant” can be used in a geographical sense to mean “most widely used,” and that may be the sense in which some people think of modern Western science as “universally valid.” (Though scientists will mean by the latter term that, for example, the laws of physics hold everywhere in the world, not just for the interactions with nature of this or that culture.) Here the term “dominant” refers, rather, to those conceptual frameworks that primarily serve the values and interests of the most powerful groups.” (Harding, 2015, p.34)

What Harding’s analysis shows is that dominant practices in scientific research, which also apply to the field of AI, have been influenced by external interests, factors, and traditions. These influences go from economic, political, and social aspects, to core methodological practices, where problems about bias, for example (as discussed in sections 1.2 and 1.3) can be narrowed to formalisations of fairness or technocentric expectations for bias mitigation. The call Harding makes, however, is not to get:

> “[…] completely outside of one’s socialization into a research discipline”, but instead understand that “finding or creating even just a little distance from prevailing assumptions and interests can be sufficient to enable critical perspective to illuminate issues in new ways.” (Harding, 2015, p. 35)

Hence, based on this, Draude et al.’s sociotechnical systemic approach to bias, inspired by these feminist theorists, requires a shift from the focus of presumed objectivity to one that is actively aware of and responsive to these power dynamics in which AI systems unravel.

In the case of technologies, more precisely AI systems, Harding’s concept of strong objectivity relates to identifying certain stakeholders, particularly those negatively impacted by AI systems, engaging with existing inequalities and analysing who benefits or is impacted by power imbalances —which includes questioning existing tendencies and practices in AI development.

In Chapter 4, when I present and analyse the BNA and the pilot case study to test it, I will refer back to this core idea related to Harding's strong objectivity, establishing that creating distance from prevailing assumptions, is a starting point to adopt much needed critical perspectives —it is a core initial step. In Chapter 4, I will argue that the BNA promotes this. One of the findings from the case study will show the reflection of the AI developers about their own professional biases and the limitations of performing under a "microscopic vision" (see section 4.3.2), which can limit their engagement with critical perspectives, that can be enabled by a strong objectivity standpoint.

The combination of these two concepts, situated knowledges and strong objectivity, grounds the systemic approach proposed by Draude et al.:

> "[…] we suggest the perspectives of situated knowledges and standpoint theory point to understanding knowledge as a product of a complex network, where human researchers, data, data structures, algorithms and broader social, political, historical and scientific contexts all contribute to the specific results that are produced." (Draude et al., 2019, 334)

Taking this into practice, the call Draude et al. make is to situate contextual factors systematically:

> "[T]o produce less biased and more accountable sociotechnical solutions, it is crucial to situate algorithmic systems and their design process, i.e. to understand and address their embeddedness in political, socio-cultural contexts and existing power structures." (p.335)

Accordingly, addressing these issues requires asking who benefits and who might be at a disadvantage, where the data originates, and the power dynamics at play.

Thus, Draude et al.'s notion of situated knowledges adapted to algorithm design is about ensuring that all aspects of the development process consider the varied impacts on different groups, involving critical questions about the origins and implications of data aiming to address the potential amplification of structural inequalities through algorithms. To achieve this, the authors suggest implementing a set of 4P questions "systematically and iteratively" (Draude et al., 2019, p.335):

- People: who are the people affected and involved? Who is benefiting?

- Place: where are the data coming from? What is contained in data and how they have been collected? How will the system affect the sources of data, if at all?

- Power: what are power hierarchies between the initiating parties, the benefiting parties and others that will be affected by the algorithmic system? Should/could these power relations be made more equal?

- Participation: who participates in the design? What kind of technological, social and cultural systems will play a role in the application of the system?

Draude et al. (2019), conclude that this necessary transition into a systemic sociotechnical approach necessitates true interdisciplinarity, recognising that their recommendations "may not be instantly actionable, [however] they serve as a navigational aid for crafting less biased sociotechnical systems" (p. 337) —calling for future work to engage in creating more tangible approaches.

This proposal of a sociotechnical systemic approach to bias gives conceptual grounding to set my Bias Network Approach proposal. The point of view from feminist epistemology and gender studies offers a direct way to translate sociotechnical views of AI ethics into practices that integrate a broader context.

The interest aspect of integrating this broader context, is that it can be more easily adapted to a promote a change of mindset within AI developers. Without being overly demanding, this conceptual take to rethinking AI bias will emphasise some of the core elements present in the other views and criticisms revised here. It highlights the need for interdisciplinarity, the relevance of inclusive design —in this case from a view of power imbalances and epistemic justice, the continuous engagement in an iterative reflective process throughout AI development, as well as being critical of technocentric and androcentric approaches to creating AI.

As Draude et al. point out, the limitation of this proposal lies in the practical application of this sociotechnical approach, a hurdle that is not unique to this framework but is common across all the perspectives reviewed in this chapter. However, it does establish a foundation for future work, considering the "increasing need for interdisciplinary methodologies, methods and tools connecting critical knowledge from the humanities and social sciences to computing." (Draude et al., 2019, p. 337)

Hence, the proposal of a BNA will respond to the need for interdisciplinary approaches, aiding AI developers in achieving this much-needed broader spectrum of analysis and contributing with an initial step to bridge this gap.

Thus, I recognise that sociotechnical approaches in AI ethics generally have a core element in common: a holistic view that recognises the sociotechnical nature of AI systems, providing a broader context to AI ethics by focusing on the relationship between technology and social structures. To complement these efforts, in upcoming Chapter 4, I will present the BNA to integrate this conceptual background into developers' practice. However, before this, there is still one conceptual discussion missing, how should we understand bias?

So far, I have presented the "problems of bias," which include the technocentric tendencies in AI, the conceptual ambiguity of the unfairness and bias conflation in the field, and the tendency to conceptualise and mitigate biases in isolation. I have also argued against this technosolutionism by examining views from various researchers that adopt sociotechnical approaches to AI ethics, emphasising the need for interdisciplinary interventions that embrace and deal with the complexities of concepts like bias.

But, before proposing the BNA to aid AI developers, I should also clarify how I think the concept of bias should be understood and defined. Considering the conceptual ambiguity mentioned, and the overly technical responses to deal with bias discussed in Chapter 1, it is important to untangle what is the understanding of bias we should have under the sociotechnical approach I will propose. In the next chapter, I will present the definition of bias that will be used for the BNA proposal.

Chapter 3: How Should We Understand Bias?

My objective in this chapter is to clarify how biases should be defined and understood in AI ethics, particularly in the context of the Bias Network Approach (BNA).

To this end, in section 3.1, I will present the most commonly identified categories of biases in AI: technical, societal, and cognitive. These categories are generally agreed upon; however, I will point out there is a gap in examining how they influence each other, and how their interaction can impact AI development— something the BNA will contribute to.

As this chapter unfolds, I will analyse two perspectives on bias. In Section 3.2, I will first examine the conceptualisation of bias as a distortion—referred to as the "negative view"—which primarily associates bias with unfairness, discrimination, or ethical failure. This perspective is deeply embedded in philosophical and AI ethics discussions, where bias is often seen as a mechanism that reinforces existing inequalities and distorts fair decision-making processes.

Next, I will explore an alternative conceptualisation of bias as a neutral and potentially positive phenomenon—referred to as the "neutral view." This perspective, grounded in epistemological foundations, suggests that biases can function as adaptive heuristics, enabling efficiency in decision-making and problem-solving. I will illustrate how this perspective has influenced AI research, particularly in discussions surrounding predictive processing models and machine learning heuristics.

Finally, in section 3.3, I will argue for caution when addressing bias within AI ethics, as our epistemic stance on bias significantly shapes the way we conceptualise and contextualise ethical issues related to bias. The way we approach bias influences how we identify, interpret, and address the various forms of bias within the AI ecosystem—a crucial consideration for the development of the BNA as an intervention aimed at fostering ethical reflection among AI developers.

Accordingly, this chapter will conclude with the definition of bias that will serve as the foundation for the BNA implementation, emphasising the importance of carefully framing discussions around bias and articulate the advantages of adopting a negative conceptualisation of bias—at least within the context of the BNA's application.

3.1 Categories for bias.

Bias in artificial intelligence is often viewed as systematic discrepancies in AI systems that lead to prejudice against specific individuals or groups, as outlined by Mehrabi et al. (2019) and Ntoutsi et al. (2020). The origin of biases in AI can be traced to multiple factors, including the way data is collected, the architecture of algorithms, and the influence of human interaction within the AI development process.

Prior literature often classifies AI biases into three broad categories: societal, technical, and cognitive. For example, for societal biases most researchers emphasise how AI systems may inherit the prejudices existing within human-generated data, leading to discrimination against certain groups as highlighted in the works of Ntoutsi et al. (2020), Ferrara (2023), and Roselli et al. (2019).

Such biases mirror and perpetuate existing societal inequalities, a situation Zajko (2022) regards as undesirable. Technical biases specifically relate to the influence of the AI development process on the algorithm's performance, such as the introduction of single-source bias when data is derived from a homogeneous system (see Rajpurkar et al. 2022).

Cognitive bias, as discussed by Soleimani et al. (2021), refers to the replication of human errors in judgment or reasoning by AI systems, arising through the interactions with their developers and users, the data they are trained on, or the inherent design of the AI algorithms themselves.

More recently, Schwartz et al. (2022) have built up on these categorisations to highlight certain sociotechnical interactions. The authors suggest that defining and describing how systemic and human biases present within AI, can allow us to build new approaches for analysing, managing, and mitigating bias and begin to understand how these biases might interact with each other.

Thus, the origin of systemic biases, according to Schwartz et al. (2022) is rooted in historical, institutional, and societal dynamics, often reflecting entrenched patterns of behaviour and structural inequalities:

> "Systemic biases result from procedures and practices of particular institutions that operate in ways which result in certain social groups being advantaged or favored and others being disadvantaged or devalued. […] Systemic bias is also referred to as institutional or historical bias. These biases are present in the datasets used in AI, and the institutional norms, practices, and processes across the AI lifecycle and in broader culture and society." (Schwartz et al., 2022, p.6)

Human biases include those that stem from unconscious errors in judgment, e.g., implicit biases, which influence decisions made during the AI development process:

> "Human biases reflect systematic errors in human thought […]. These biases are often implicit and tend to relate to how an individual or group perceives information (such as automated AI output) to make a decision or fill in missing or unknown information. These biases are omnipresent in the institutional, group, and individual decision making processes across the AI lifecycle." (Schwartz et al., 2022, p.9)

And finally, statistical and computational biases, pertain to technical imbalances in representation and systematic errors in data processing, which manifest as biases in the statistical algorithms and can lead to observable favouritism or discriminatory outcomes in AI systems:

> "Statistical and computational biases stem from errors that result when the sample is not representative of the population. These biases arise from systematic as opposed to random error and can occur in the absence of prejudice, partiality, or discriminatory intent. In AI systems, these biases are present in the datasets and algorithmic processes used in the development of AI applications and often arise when algorithms are trained on one type of data and cannot extrapolate beyond those data." (Schwartz et al., 2022, p.9)

What is interesting about Schwartz et al.'s categorisation, is that it presents these categories suggesting that they are somehow linked or connected (as shown in Figure 4).

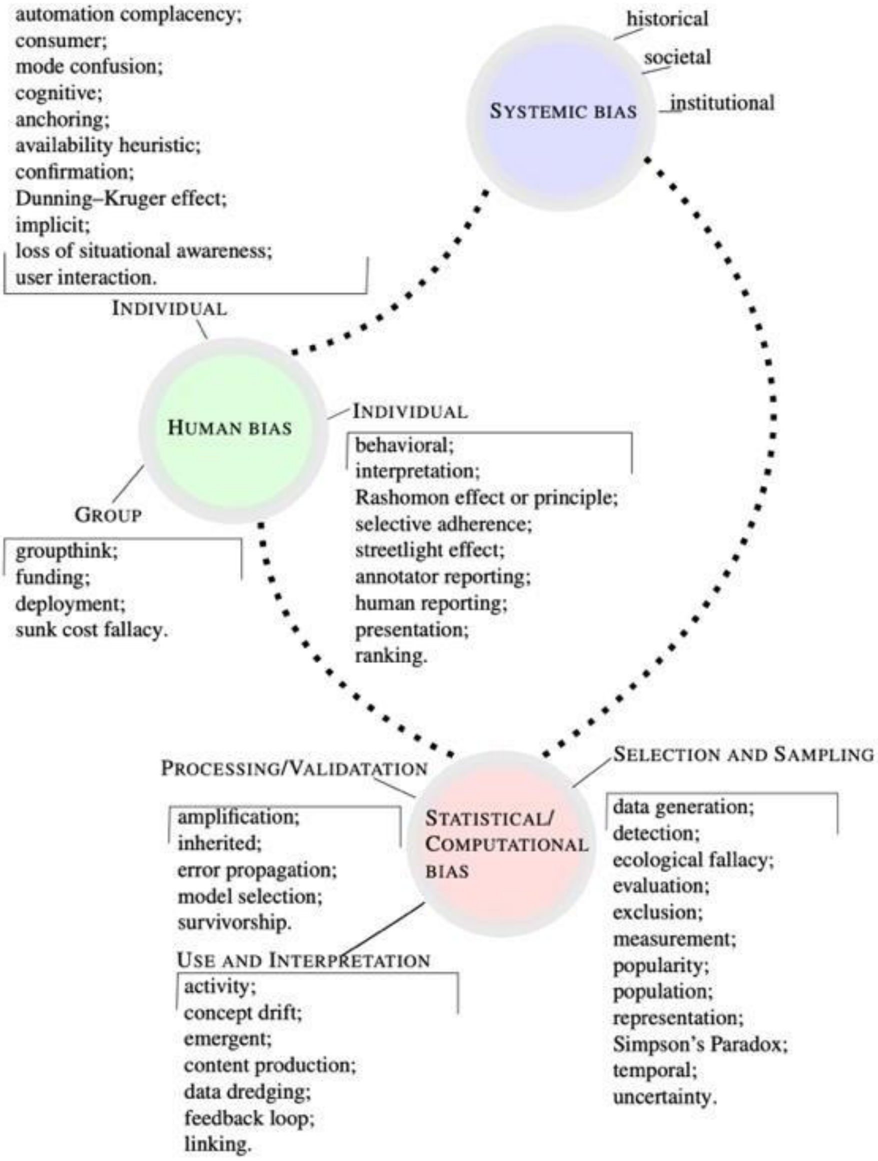


Figure 3: Categories of biases by Schwartz et al. (2022).

Although they do not engage with this any further, they do emphasise the inherent sociotechnical aspect of challenges derived from AI biases.

For example, the authors discuss the issue with proxies. Even when datasets are deemed representative, they can still embed historical and systemic biases, misuse protected attributes or employ culturally and contextually unsuitable attributes. To avoid biases, some developers may exclude protected attributes linked to historically discriminated social groups.

Nevertheless, this approach may not effectively address the problem, as the information can be inferred through proxy or latent variables. Furthermore, biases can also arise from user behaviour and feedback loops. The presence of feedback loops can lead to disparity amplification within AI systems, wherein marginalised individuals or groups are less inclined to use the technology, resulting in the subsequent training data predominantly reflecting the behaviours of the most frequent users.

A feedback loop in AI occurs when the model's predictions influence the data it will learn from in the future. This can cause the model to reinforce its own biases. Imagine a recommendation system that starts prioritising a suggestion for pop music. The more it recommends pop music, the more users listen to it, and the more the system identifies pop as the preferred genre, it recommends more pop music. Over time, this loop can make the system heavily biased toward pop, even if users would enjoy a wider variety of music. However, these biases become even more problematic when they create a harmful feedback loop that ostracises users.

For example, studies have shown that voice-enabled assistants work consistently better for native English speakers (Song et al., 2022; Zwakman et al., 2021), which makes non-native speakers less likely to use them. As a result, the data collected by these systems will mostly come from native speakers, reinforcing the bias because AI is not exposed to diverse accents or speech patterns, so it does not learn to understand them better.

Consequently, the experiences of these specific groups do not align with the intended purpose or functioning of the AI system. This phenomenon perpetuates disparities and highlights the potential for biased outcomes in AI applications. Another bias-related issue discussed by Schwartz et al. is the epistemic uncertainty that arises in deep learning models due to the non-unique solutions obtained during the nonconvex minimisation of the cost function that is used to compute model parameters.

At the heart of a machine learning algorithm, is the cost function, also known as the loss function. This mathematical function measures the error between the model's predictions and the actual observed outcomes. The central task during the training phase is to adjust the model parameters in such a way that this error is minimised. For a model to be considered accurate and reliable, the values predicted by the model must closely align with real-world data.

Deep learning models, which include neural networks (NN), rely on complex architectures representing nonlinear relationships. Because of this complexity, the cost function landscape is nonconvex, i.e., the surface described by the function has multiple valleys or local minima—points where the function value is lower than in the immediate vicinity, although not necessarily the lowest.

The epistemic uncertainty issue refers to the lack of knowledge about the process that generated the data. This uncertainty arises because of the nonconvex nature of the cost function, which may yield numerous potential solutions during the optimisation process. Each set of parameters that corresponds to a local minimum could serve as a potential solution to the optimisation problem, but not all solutions are equally valid or accurate in terms of real-world performance.

The practical implication of this complex mathematical scenario is that it introduces a level of uncertainty into the model's predictions, which can result in optimisation results that are less accurate or generalisable to new, unseen data. This is particularly problematic when the model is applied in critical domains where errors can have significant consequences.

In simpler terms, when training a deep learning model like a NN, we try to adjust it so that its predictions are as accurate as possible to the real world, i.e., we "tune" a complex machine to get the best performance we can. The tool we use for this tuning is the cost function, and it measures how far off the model's predictions are from what they should be (or at least what we have established as an acceptable parameter). Hence, the goal is to make adjustments that lower this cost as much as possible, which improves the accuracy of the model. However, deep learning models are very complex and can be tuned in many different ways.

Imagine you are trying to find the lowest point of a mountain landscape at night to get out of a park. If the landscape had a convex shape, you could just walk straight to the bottom and find your way out. The problem is that NN do not have a convex landscape, and instead, theirs looks more similar to a set of valleys, a nonconvex shape. As you walk towards one of these valleys, you might think you found the lowest point, but there could be an even lower one you do not know about, making it easier to get out of the park. Each time you walk around trying to find the lowest point (train the model), you might end up in a different valley (solution). This is epistemic uncertainty.

This uncertainty can lead to different kinds of biases in models. For example, if you always start your search from the same spot (maybe your favourite lookout), you might keep finding the same valley even if it is not the lowest. Similarly, a model can continue to learn the same patterns from the data, even if they are not the best options available. Or perhaps a trickier example, is when the model finds valleys that are quite good for certain types of data but completely wrong for others. This would mean that it might have a good general performance but have a strong bias against certain types of data, which can translate into a bias against certain groups, individuals, or contexts.

This translates to unreliable and even unfair predictions when the model is used. Increasing representative training data and adopting a bias-aware approach could mitigate epistemic uncertainty in certain cases, but the complete elimination of these biases remains unattainable.

With these examples, I wanted to show the variety of origins and categories of biases that can be actively influencing AI's development process and the developers' perspective. Therefore, if the biases reflect societal issues in the data, if they are part of the cognitive background of the developers, or if they are present in the model training stage, they should be understood in relation to their context, as suggested in my criticism against isolationism in section 1.3.

3.2 Two views on bias: negative versus neutral.

In this section, I will introduce two ways in which bias is defined and studied in philosophy, both of which have influenced how bias is understood in AI. First, I will examine the "negative" view on bias, that see it as an undesirable phenomenon, primarily linking it to instances of unfairness or discrimination in ethical discourse. I will illustrate how this view has permeated discussions in AI ethics.

Next, I will examine views on bias as a "neutral" or potentially advantageous phenomenon, primarily grounded in epistemological foundations. I will provide examples of how this view has been incorporated into AI research, highlighting its implications for discussions on bias.

3.2.1 Bias as negative

The term bias originated in the 1500s, first used in lawn bowling to describe a built-in weight imbalance that caused the ball to curve off a straight path (Kelly, 2022) and as noted by the *Oxford English Dictionary* (OED)*,* "bias" started in the English language as a noun for an "oblique or slanting line." The earliest *OED* example for bias is from the French Grammar book *Lesclarcissement de la Langue Francoyse* (1530) by John Palsgrave, used as reference to going against the grain or sideways. With time, it has evolved into what we understand today, keeping that original idea of "disproportionate weight."

This idea of bias as a force that distorts movement or decision-making persists today, shaping discussions in fields like cognition, ethics, and artificial intelligence. Thomas Kelly (2022), in his study about bias, recognises both the negative and the neutral accounts for bias. For the first one, he recognises that this understating of bias as morally undesirable is linked to a norm-theoretic account of bias: "a bias involves a systematic departure from a norm or standard of correctness." (Kelly, 2022, p. 63) Moreover, he mentions that bias is frequently framed in a negative or pejorative sense, particularly when it is understood as a systematic deviation from a moral or procedural norm.

Unlike random errors or inconsistencies, biased beliefs, decisions, or processes exhibit structured tendencies that deviate from what is the norm, i.e., a distortion. These deviations are not merely accidental but reflect a persistent pattern that leads to unfair, inaccurate, or misleading outcomes.

The systematic nature of bias, in this sense, makes it particularly problematic, as it can distort perception and decision-making in ways that are resistant to correction, and even in cases where there might be more fundamental failings, biases can still be regarded as morally significant and, therefore, undesirable:

> "When an agent is biased in the pejorative sense, they are typically guilty of some other failing or shortcoming that in some respects is more fundamental. However, this need not diminish the significance (moral or otherwise) of the fact that the agent is biased as opposed to merely guilty of the more fundamental failure or shortcoming. Indeed, it's perfectly consistent with what's been argued here that the bias and the failure to which it leads are morally significant, even if the characteristic failure that is its manifestation would not be, if that mistake had occurred as a result of random error." (Kelly, 2022, p.104)

Kelly's account contrasts bias as norm-violating with instances where bias may be present but not necessarily problematic. The critical distinction lies in whether the bias causes an ethical failure. For example, a person who consistently interprets evidence in a manner that unjustifiably favours their preexisting beliefs would be considered biased in a negative sense because their reasoning deviates from an epistemic norm of objectivity and truth-seeking. Similarly, a legal system that consistently favours a particular demographic group in sentencing decisions would exemplify a morally problematic bias due to its departure from norms of justice and impartiality.

Hence, in its negative sense, bias hinders rational deliberation and ethical decision-making, reinforcing patterns of distortion that perpetuate injustice, misinformation, and irrationality. This negative view of bias can be applied to a broad range of entities, as noted by Kelly (2022, pp. 24-25).

Individuals, particularly in social roles such as judges or committee members, are frequently characterised as biased. Bias is attributed to inanimate objects, such as a loaded dice, as well as to sources of evidence and information, including surveys, research studies, and media outlets. The concept extends further to procedural mechanisms, such as hiring and admissions processes, as well as cognitive and linguistic phenomena, including beliefs, narratives, and texts. Kelly also mentions how in contemporary discussions, biased algorithms have also become a focal point of concerns.

Ordinarily, attributing bias carries a negative connotation, often implying a failure of objectivity or fairness.[8] Accusations of bias in judicial rulings, historical interpretations, scientific studies, or political polling generally suggest that the judgments or findings in question are unreliable or compromised. This normative aspect of bias attribution highlights its function not just as a descriptive claim but as a form of critique, challenging the legitimacy of particular decisions, interpretations, or processes.

Within this negative framework, one particularly scrutinised type of bias in philosophy—due to its moral undesirability—is implicit bias. Ethical discussion about implicit bias generally focuses on how it distorts our social behaviour and judgment.

Jennifer Saul, argues that the most worrying aspect of implicit biases is that they manifest as subconscious, automatic inclinations to associate specific characteristics with certain social groups, leading to serious mistakes:

> "[…] unconscious tendencies to automatically associate concepts with one another. Put like this, they don't sound very interesting or worrying. But the ones on which attention by philosophers has focused are both very interesting and very worrying. These are unconscious, automatic tendencies to associate certain traits with members

[8] It is worth making a clarification here. In his work, Kelly identifies a key difference, that also contributes to the clarification of bias I wish to develop in this Chapter and that follows some of the issues about bias being conflated with fairness in Chapter 1. The distinction between bias and unbiasedness is more nuanced than a simple binary opposition. The claim that something is *not biased* does not automatically mean that it is *unbiased*, a distinction that is essential to understanding how bias functions within epistemic and social contexts. In particular, biased and unbiased are best understood as *contraries* rather than *contradictories*. This means that while something that is unbiased necessarily implies that it is not biased, the reverse does not hold—something that is not biased does not necessarily qualify as unbiased. This distinction complements the previous discussion where I argued that it is more accurate to say that bias is often conflated with unfairness. On top of this, we should be careful in not equating unbiased with not biased. Many things that lack bias do not necessarily embody unbiasedness. A rock, a desk chair, or the number 17, for example, cannot be meaningfully described as biased, but neither do they count as unbiased in the relevant sense. This highlights the fact that the property of unbiasedness presupposes the potential for bias—only entities capable of demonstrating bias can meaningfully be described as unbiased.

> of particular social groups, in ways that lead to some very disturbing errors." (Saul, 2013, p.244)

Ultimately, Saul argues that bias is undesirable because it leads to morally and politically disturbing consequences: "dramatically unfair in our judgements, even though we are doing it unintentionally." (Saul, 2013, p. 246)

According to Saul, biases obscure judgment and performance by allowing irrelevant factors to influence assessments. More specifically, she argues that we frequently make errors by permitting an individual's or group's social identity to shape our evaluations—despite believing that such considerations should not affect our reasoning. This aligns with the broader negative view of bias, where bias attribution is tied to a normative failure. As Kelly (2022) notes:

> "[…] when someone claims that a particular interpretation of an historical event or a text is biased, we naturally take them to be disputing that interpretation; someone who claims that a scientific study or political poll is biased is naturally understood as suggesting that its putative findings shouldn't be accepted at face value" (p.25)

Similarly, Sally Haslanger (2015) conceptualises biases as cognitive structures or schemas that shape perception, thought, and action. She highlights the insidious nature of biases, which often operate subtly and remain unrecognised, manifesting through internalised behaviours. Haslanger situates biases within broader social structures and systemic injustices, arguing that they not only distort cognitive schemas but also perpetuate and exacerbate deeply entrenched inequalities. From this perspective, biases are not merely individual failings but integral components of social systems with far-reaching consequences. Haslanger further emphasises that addressing biases requires deliberate effort, as their identification and mitigation demand conscious reflection and active intervention.

Consequently, biases are not merely individual failings but integral components of social systems with extensive and often detrimental implications. Haslanger's analysis focuses on how biases are problematic, because we need a deliberate effort to recognise, manage, and mitigate their impact—a challenge that confronts us all:

> “because there is empirical evidence to support the claim that we are all biased; insofar as we are able to control or change our biases, it is a potential site for moral responsibility and moral improvement” (Haslanger, 2015, p.12)

Overall, under these types of views, biases are negative because they influence our choices and actions, which are not based on a fair assessment but on preconceived notions or unjustified judgements, leading us to deviate from morally desirable norms.

In the field of AI, the negative implications of bias are also widely acknowledged, primarily in discussions surrounding the systematic unfairness of AI systems' outcomes. As illustrated in Section 3.1, this unfairness can arise from multiple sources, including the data used to train the system, the design of the algorithm, and the broader societal and historical context in which the technology operates. These biases are not merely incidental flaws but are often deeply embedded in the development, deployment, and interaction of AI systems with human users.

Moreover, if we adopt a sociotechnical perspective on AI, as I do in this work, AI bias extends beyond a machine making isolated errors. Instead, it reflects a pattern of systemic distortions that disproportionately disadvantage certain groups or individuals. These biases reinforce existing social inequalities and can manifest through complex interactions between AI systems and human decision-making.

Thus, AI bias is not merely a technical issue but a broader structural concern, necessitating critical engagement with both its computational underpinnings and its ethical and social implications.

Mehrabi et al. (2021), for example, frame the notion of “unfair algorithms” based on their vulnerability to biases:

> “[L]ike people, algorithms are vulnerable to biases that render their decisions “unfair”. In the context of decision-making, fairness is the absence of any prejudice or favoritism toward an individual or group based on their inherent or acquired characteristics. Thus, an unfair algorithm is one whose decisions are skewed toward a particular group of people.” (Mehrabi et al., 2021, p.1)

This way of thinking about bias, however, recalls my criticism of the bias-centric view of fairness, as it highlights that fairness is the absence of prejudice or biases.

Other researchers like Hellström et al. (2020), emphasise the normative and legal meaning of bias —which also coincided with our use of the concept of bias in ordinary language:

> "The word 'bias' has an established normative meaning in legal language, where it refers to 'judgement based on preconceived notions or prejudices, as opposed to the impartial evaluation of facts.' The world around us is often described as biased in this sense, and since most machine learning techniques simply mimic large amounts of observations of the world, it should come as no surprise that the resulting systems also express the same bias." (Hellström et al., 2020, p.2)

Ultimately, the ethical worries about the presence of AI bias, are most commonly rooted in the different ways in which bias leads to instances of discrimination, frequently present in the form of "algorithmic bias", as Chen describes:

> "The algorithms frequently contain […] biases due to the lengthy history of racial and gender prejudices, both intentional and unconscious. When biases exist in algorithmic data, AI may replicate these prejudices in its decision- making, a mistake known as algorithmic bias." (Chen, 2023, p.5)

And, as Ferrara points out, addressing biases becomes relevant to ensure that AI is fair to *all* users, thus avoiding discriminatory and unfair outcomes:

> "Bias is defined as a systematic error in decision-making processes that results in unfair outcomes. In the context of AI, bias can arise from various sources, including data collection, algorithm design, and human interpretation. Machine learning models, a type of AI system, can learn and replicate patterns of bias present in the data used to train them, resulting in unfair or discriminatory outcomes. It is important to identify and address bias in AI to ensure that these systems are fair and equitable for all users." (Ferrara, 2023, p.2)

Despite the different descriptions, all these references show a negativity related to how they conceptualise and study bias. On the negative view, when we claim something or someone is biased, we mean something has gone wrong, particularly in relation to human judgement or behaviour.

Hence, adopting this view puts us in a particular epistemic stance, a state of alert where we want to avoid or address biases. Even when we acknowledge that it is not possible to get rid of all biases completely, we want to at least identify their influence or impact because it creates (directly or indirectly) a harmful effect.

3.2.2 Bias as neutral.

Unlike the previous view, an alternative perspective on bias challenges the assumption that biases are negative. Some philosophers argue that biases can be beneficial, even essential, for cognitive processes.

For example, Andy Clark explores the role of biases within his predictive processing framework, which conceptualises the brain as a dynamic prediction engine (Clark, 2016; 2024). This framework suggests that cognition is not a passive process of receiving sensory input but an active process of generating and refining models to anticipate and interpret the world, thereby minimising prediction errors.

Within this predictive system, biases function as adaptive cognitive shortcuts, allowing the brain to efficiently filter vast amounts of information and prioritise what is most relevant. Rather than being mere distortions, biases serve as heuristic mechanisms that optimise cognitive efficiency by guiding attention and shaping expectations. These biases help structure perceptual and inferential processes, enabling individuals to navigate complex environments with limited cognitive resources.

To fully grasp this perspective, it is useful to consider Clark's definition of the predictive brain:

> "The predictive brain, if this is correct, is not an insulated inference engine so much as an action-oriented engagement machine. It is an engagement-machine, moreover, that is perfectly positioned to select frugal, action-based routines that reduce the demands on neural processing and deliver fast, fluent forms of adaptive success." (Clark, 2016, p.1)

Biases operate within this Bayesian-inspired model of cognition. By understanding biases as integral to predictive processing, this framework presents a fundamentally different account of bias, one that recognises its role in shaping effective and adaptive cognition rather than solely as a source of epistemic or moral failure:

> "Sub-cortical influences here bias large-scale neural patterns towards signals that are *biologically valuable* – those accorded high precision within the PP scheme." (Clark, 2018)

Clark does not give a specific overarching definition to bias in his work and instead uses it on most cases to exemplify a series of adaptive shortcuts that happen within his perceptual framework.

For example:

> "Within the PP framework, gating is principally achieved by the manipulation of the precision-weighting assigned to specific prediction errors. The primary effect of this […] is to systematically vary the relative influence of top-down versus bottom-up information by increasing the gain ("volume') on selected error units. This provides a way to implement a rich set of attentional mechanisms whose role is to bias processing so as to reflect estimates of the reliability and salience of (different aspects of) both the sensory signal and the generative model itself'." (Clark, 2016)

This illustrates how biases emerge as a consequence of gating mechanisms that regulate the balance between top-down expectations and bottom-up sensory input. This is achieved through precision-weighting, where the brain selectively amplifies or suppresses prediction errors based on their estimated reliability. By increasing the gain ('volume') on certain error units, the system biases processing toward either reinforcing prior beliefs or prioritising new sensory data. This adaptive mechanism underlies cognitive biases such as perceptual biases, where unreliable sensory input is discounted in favor of a stable, internally generated model. Attentional biases also emerge from this process, as the system selectively enhances information deemed salient or reliable, shaping what is perceived and acted upon, thus optimising cognitive efficiency.

Furthermore, he also describes biases as inbuilt in our evolved cognitive structure:

> "Of course, as King Lear famously commented, 'nothing will come of nothing', and, as hinted above, even the most slimline learning system must always start with some set of biases. More important, our basic evolved structure (gross neuroanatomy, bodily morphology, etc.) may itself be regarded as a particularly concrete set of inbuilt (embodied) biases that form part of our overall 'model" of the world" (Clark, 2016, p. 175)

Rather than being mere cognitive flaws, for Clark, biases emerge as indispensable tools for navigating complex and uncertain environments. For instance, biases such as confirmation bias or anchoring bias reflect the brain's reliance on prior expectations to streamline decision-making, reducing the cognitive load required to evaluate every possibility afresh. These biases are particularly useful in situations where time or information is limited, as they allow individuals to make rapid, often effective judgments:

> “For wherever prediction helps construct experience there is a kind of bias. The world as we see and sense it becomes shaped, in part, by our own (conscious and unconscious) expectations. This is not merely bias in thought or judgment but bias affecting the primary sensory realm—the source of our apparent evidence—itself.” (Clark, 2024, p.117)

Clark further emphasises the context-sensitive and embodied nature of biases, highlighting their role in shaping perception and action through interactions with the physical and social environment. He argues that biases are not fixed errors but flexible, context-dependent tools that adapt to specific situations.

This flexibility is central to the brain’s capacity to learn and recalibrate its predictions when confronted with persistent errors. Using the metaphor of a surfer navigating waves, Clark illustrates how biases allow individuals to maintain balance and trajectory by leveraging pre-learned patterns and environmental cues. This view emphasises that although biases can occasionally lead to errors, such as perpetuating stereotypes or cognitive distortions, they also play a vital role in ensuring cognitive efficiency and adaptability. Clark’s work thus reframes biases as essential components of human cognition, offering a nuanced understanding that bridges theoretical insights with practical implications for fields such as artificial intelligence, decision-making, and behavioural sciences.

Another example of this neutral view on bias is the idea proposed by Louise Antony. For her, bias can be defined as an inclination of temperament or outlook (Antony, 2016) framing bias as a necessary and constructive component of human cognition. Based on Quine’s view, Antony argues that empirical learning could not proceed without having an innate ‘similarity space’, thus biases make salient certain properties of experienced objects. This is cognitively beneficial and necessary, as it is part of what we have developed to acquire empirical knowledge.

Antony follows a naturalistic method, responding to Saul's sceptical view: "Saulish scepticism is a phenomenon that shows us the need to take a naturalized approach […] [which] reveals that bias is an essential and constructive factor of our ability to know the world" (Antony, 2016, p. 188). According to Antony, the ability to group and differentiate stimuli based on perceived similarities is an innate mechanism that allows for the structuring of hypothesis spaces and the simplification of complex data. Without such predispositions, empirical inquiry would become intractable, overwhelmed by the sheer volume of possible interpretations of sensory inputs. Hence, she claims that taking this stance helps understanding how and when bias is friend or foe; bias is not something we ought to consider bad a priori but rather realise when biases can become troublesome.

Antony's defence of bias, therefore, extends to its role in aligning cognitive resources with situational demands. A fundamental implication of the naturalistic approach is the necessity of reconsidering a certain conception of objectivity. A naturalised perspective on human knowledge challenges the prevailing assumption that epistemic success requires the eradication of all bias. Within this framework, objectivity—understood as the complete absence of bias—emerges as an untenable epistemic ideal. Bias, far from being merely an inevitable feature of human cognition, constitutes a fundamental enabler of epistemic achievement. It is not merely an obstacle to be overcome but rather an essential component of the processes through which knowledge is constructed and refined.

In this sense, Antony argues that bias plays a constructive role in the acquisition and development of knowledge, i.e., presents a cognitive advantage. It serves as an enabling condition that structures cognitive processes, allowing individuals to navigate vast and complex informational landscapes.

Rather than impairing epistemic success, bias facilitates it by helping to organise, prioritise, and interpret empirical data in ways that render learning and inquiry manageable. In the absence of bias, the sheer multiplicity of potential interpretations would overwhelm cognitive faculties, thereby obstructing rather than enhancing our capacity to understand the world. She contends that biases are not problematic but become so only when they lead individuals away from truth or justice. Hence, biases are foundational to how humans acquire and process empirical knowledge. In such cases, biases serve as cognitive shortcuts, enhancing decision-making in high-pressure environments. This is why she defines it as a *tendency*:

> "The term "bias," as it is commonly used, implies something morally or rationally negative. I mean to use the term in its more general, normatively neutral sense, as meaning "a tendency; an inclination of temperament or outlook."" (Antony, 2016, p.161)

Antony's view reframes bias as not merely a source of distortion but as a vital cognitive tool that, when properly understood and managed, contributes to epistemic success and adaptive functioning. Hence, the neutrality in its definition.

If one accepts this and considers that bias is not merely a cognitive hindrance but an epistemic necessity, then calls for its elimination fail to account for its dual role in human cognition. Instead, what Antony hints at with the idea of bias being "friend or foe" is to adopt a more nuanced approach—one that seeks to distinguish between biases that contribute to epistemic success and those that distort or obstruct it.

Recognising this complexity reframes the ethical and epistemological challenges posed by bias, compelling a more sophisticated engagement with its implications in both theoretical and applied contexts. Antony offers a possible approach to engage with recognising when bias is friend or foe, that is, by understanding the origins and functions of constructive biases, we can better identify and mitigate those that are harmful or epistemically pernicious.

To determine whether a bias is constructive or pernicious, Antony suggests examining its ecological validity—that is, whether the markers we use in judgment accurately correspond to the target properties we aim to track. The proposed framework consists of three conditions: (i) markers and targets – what properties are we trying to identify (targets), and what observable features do we use as indicators (markers) of those properties; (ii) indication relations – clarify if these markers reliably correlate with the intended target properties; and (iii) sustaining mechanism – if a correlation exists, what mechanism underpins and maintains it.

A pernicious bias, according to Antony, is one where a practice fails the second condition, meaning the markers being used do not actually correspond to the target properties. She presents the following example to illustrate this:

> "Consider epistemic authority: if there has been a long period of time in which intellectual experts were almost all men, then a deep voice might have become a marker of expertise. If it is no longer the case that intellectual experts are almost always men—if there have come to be a significant number of women who are entitled to be treated as authorities, then the "deep voice" marker has gotten out of tune with the trait it once marked. We will, in this sort of case, need to contrive ways of deactivating that marker and, eventually, replacing it with one that is more reliable." (Antony, 2016, p.184)

Thus, Antony's view tells us that instead of aiming to eradicate bias entirely, we can resort to a more pragmatic approach: manipulating the environment so that biases can function in ways that do not perpetuate injustice. This perspective builds on the idea that bias is an inherent and often necessary part of cognition. Biases help us make sense of the world, filter information, and navigate complex environments. The problem arises when biases systematically distort judgment in ways that reinforce inequality or exclusion. For example, if gender bias leads people to associate authority with deep voices, the solution is not merely to suppress this bias but to ensure that authority is distributed equitably, so that deep voices are not the only voices associated with leadership.

Antony offers a shift in focus: rather than treating bias as a moral failing or an epistemic defect that must be eradicated, we should acknowledge its role in cognition and engineer our social institutions to mitigate its harmful effects because as she stresses: "gaining an understanding of how and when bias is our friend will enable us to act more effectively when it becomes our foe." (Antony, 2016, p.188).

Scholars working in AI, have also suggested neutral accounts of bias that, similarly to the ones mentioned above, emphasise a positive impact or advantageous presence of biases, for the learning and training aspects of AI systems.

Pot et al. (2021), for example, argue that “not all biases are bad” (p.8) challenging the view of bias as negative, by carefully analysing the role of bias within the context of ML in medicine, particularly in radiology. The authors argue that in ML, biases are not only technical issues requiring technical solutions; they have a social dimension that can impact equity in healthcare outcomes:

> “For example, the composition of patient dataset used for research in imaging (e.g. population imaging) is influenced by who has access to radiology services in the first place. […] The people who do have the best access to radiology services are most likely the ones most benefitting from the application of the ML technology, because they were represented in the algorithm’s training data.” (Pot et al., 2021, p.2)

Accordingly, they propose “to understand bias as, first and foremost, a social problem” (ibid), for which, one should analyse the causes and implications of biases through an equity healthcare framework. Based on this view, the authors argue that while certain biases can lead to inequities and must be addressed, other biases may help overcome existing inequalities in healthcare. Pot et al., differentiate between biases that distort outcomes in harmful ways and those that could potentially contribute to more equitable healthcare outcomes. This distinction is crucial, considering the potential of ML to either mitigate or exacerbate disparities in healthcare delivery:

> “Some biases have harmful consequences for some groups of patients and are unjust. But this does not apply to all biases: In certain cases, the creation of deliberate bias in datasets, for example, can make decisions emerging from machine learning technologies more equitable.” (Pot et al., 2021, p.1)

For example, the authors suggest that creating a deliberate bias in datasets can be beneficial. They suggest that if data from underserved populations are intentionally oversampled to compensate for their previous underrepresentation, this deliberate bias can have a beneficial effect:

> “[I]t may be necessary in some cases to create biases on purpose. For example, if data from underserved populations—such as economically deprived groups—are oversampled to compensate for a previous invisibility of these groups, then the data collection has a deliberate bias that seeks to create a beneficial effect, namely to prominently include a group that had previously been marginalised.” (Pot et al., 2021, p.6)

The authors also distinguish biases that arise in the medical context, that are part of everyday medical practice, manifesting as overt or subtle prejudices, affecting the diagnosis and the overall decision-making process of healthcare professionals. For this, actively "biasing" a training dataset, for example, is presented as a solution for which bias becomes "not bad":

> "Another example would be to deliberately oversample people with darker skin for a dataset training an algorithm detecting skin diseases as certain colour contrasts may be less easy to discern on dark skin than on light skin. In these cases, biases are explicitly equitable." (Pot et al., 2021, p.6)

Therefore, examining its nature, Pot et al. (2021) argue in favour of viewing bias as a social problem, distinguishing between harmful biases and those that might be benign or even advantageous: "instead of automatically assuming that all biases are "bad", we propose to think of some biases as "good" and desirable, because they can help to overcome existing inequities." (p.6) Overall, their main arguments point to instances in which biases can be either problematic or unproblematic:

> "We have argued that not all biases are bad: biases can be problematic and unproblematic. They are unproblematic if they contribute to greater equity […], meaning that they are based on or create a distortion of reality that is not unjust and might even be beneficial. Biases are problematic if they are inequitable. […] Biases are unjust in a distributive sense if they lead to an unfair distribution of goods such as access to healthcare services. […] From a relational justice perspective, ML algorithms are unjust if they are used for objectives that undermine equal respect and dignity among patients, independently of whether they are biased in a technical sense. Finally, biases may be relationally unjust if concerns about the use of algorithms or their outcomes are not being taken seriously and people's concerns are dismissed." (Pot et al., 2021, p.8)

Overall, Pot et al. (2021) present a re-examination of biases in machine learning, urging a departure from the view that biases are intrinsically negative. They argue that some biases, when carefully managed and directed towards addressing social inequities, can have positive effects.

Another neutral view on bias has been presented by Sara Fabi and Thilo Hagendorff (2022). They argue that intentionally including human cognitive and ethical machine biases can enhance AI systems. Their view on ethical machine biases as biases that are rooted in the deliberate selection and weighting of features in machine learning datasets to promote prosocial attributes and goals, i.e. “how the world should be”:

> “We argue that one should reintroduce the idea of algorithmic discrimination in an altered, positive manner. The basic idea is that in principle, data sets contain features that should be weighted stronger than others, perpetuating particular machine biases in an intentional manner. In this context, one can differentiate between “the world as it is” versus “the world as it should be” (Hellström et al. 2020). Models can be used to predict the world “as it is”, which means to perpetuate random existing biases. Debiasing training data, in contrast, can lead to a modeling of the world “as it should be”. Here, we also opt for using an understanding of “the world as it should be”, but, instead of debiasing, by intentionally introducing bias.” (Fabi & Hagendorff, 2022, p.13)

Instead of passively modelling the world “as it is,” which often perpetuates existing biases found in historical data, the goal is to shape the world “as it should be,” say the authors.

They discuss the necessity and potential benefits of intentionally incorporating biases into AI:

> “The idea to include cognitive biases into machine learning algorithms has, at least to our knowledge, not been raised in the literature so far, since human biases have long been seen as violations of rationality standards, as limitations to intelligence, or simply as flaws. Nowadays, a more nuanced and positive view of human cognitive biases has been established that leads to the idea of including those into machines. We argue for a re-evaluation of the notion of ethical and cognitive biases in machines, which can be ethically desirable as well as methodologically advantageous when implemented in machine learning models.” (Fabi & Hagendorff, 2022, p.2)

To support this idea, the authors argue that we can introduce deliberate biases into the data. Hagendorff and Fabi recognise that this idea “to intentionally include ethical data biases goes against the mainstream discourse in the field,” (p.2) which follows the “negative views” discussed in the previous section. Nevertheless, they propose that rethinking the role and significance of deliberately promoting and integrating biases is crucial.

Thus, the authors argue that biases can be beneficial for AI systems —in line with what Pot et al. (2021) suggest. More specifically, the authors give some specific examples of when intentionally introducing ethical machine biases and cognitive human bias. In the case of ethical machine biases, one of their examples suggests introducing representational bias in social media recommendation systems to promote rational, reflective interactions over impulsive ones.

Representational bias arises when machine learning models are trained on a dataset that favours certain representations. Fabi and Hagendorff, claim that fostering representational biases in training datasets can help avoid fake news, extremist content, and filter bubbles:

> "[…] when taking social responsibility seriously, platforms should rearrange their objectives towards values of a vital and fair public discourse, truth, and information quality. This means to change the methods for algorithmic measurement and determination of information relevance. In order to achieve this, platforms have to foster representational biases in training data sets –for instance by favoring representations of rational, effortful, reflective interactions over impulsive interactions." (Fabi & Hagendorff, 2022, p.15)

What the authors propose is that instead of just showing users content that gets a lot of likes or shares, these platforms should aim to highlight content that supports meaningful and truthful conversations. To do this, they would need to change how the algorithm decides what content is important, focusing more on posts that people spend time reading and thinking about, rather than those that are just quickly passed along. For instance, if someone takes their time to write a comment or reads through a post slowly, the algorithm should notice this and could show these kinds of posts more often.

Another example they use suggests that content production bias in language generation can be useful to influence the original text sources on the language produced by AI systems, like chatbots and speech assistants. This bias emerges from the varying qualities of user-generated content, which can range from formal, well- edited text to casual, error-filled, or simply biased language. When an AI model is trained to generate language, it learns from patterns in the data it is fed. If the training data includes a wide variety of text sources —like books, articles, and social media posts— the model will pick up on the structural, lexical, semantic, and syntactic patterns of these texts. This can be problematic if the model learns from texts that contains social biases, use poor language, or display various forms of discrimination.

For instance, if a chatbot is trained on internet forums where aggressive or discriminatory language is common, it may replicate those patterns in the output. To prevent this, the selection of training data for language generation models, say the authors, should be biased towards high-quality content:

> “The selection of corpora should be intentionally biased towards narrowing it down to digital writings that underwent a firm quality check through publishers, peer reviews, or media agencies, that are embedded in a sophisticated web of citations or links, or that stem from individuals with high levels of language skills. […] By using these selection criteria for content production, biases are purposefully implemented in natural language models. Content production biases thus are improving the quality of natural language generation.” (Fabi & Hagendorff, 2022, p.17)

Overall, their argument for introducing ethical machine bias is based on the importance of selecting and filtering training data according to ethical criteria. Regarding the introduction of cognitive human biases into AI, the authors suggest that, in certain situations, this could improve the performance and ethical decision-making of AI systems. One example they discuss to show this potential benefit is overfitting avoidance. Overfitting is an unwanted machine behaviour that happens when an ML model performs well on or accurately on training data, but not with new data.

To explain overfitting, the authors consider the following example. In human decision-making, heuristics are simple rules or mental shortcuts that focus on the most important aspects of information while ignoring the rest. This can be beneficial because it allows us to make quick decisions without getting bogged down by too much information, which might not be relevant or could even be misleading. To analyse this, they make a parallel between human heuristics and techniques used in ML to prevent overfitting:

> “According to Gigerenzer and Brighton (2009), the amount of information that humans need to ignore in successful decision-making correlates with increasing unpredictability. An analogy can be drawn nicely for artificial neural nets, which are dealing with uncertainty: The more unpredictable the data is, that is the more training and test sets differ, the greater the problem of overfitting gets, if the algorithm has too many free parameters, and thus, becomes too specialized on the training set. As described above, we claim that cognitive biases can help to avoid overfitting not only in humans’ models of the environment.” (Fabi & Hagendorff, 2022, p.10)

Methods like feature selection, where irrelevant data is removed before training, and early stopping, where training is halted before the model becomes too specialised, mirror the human approach of ignoring less relevant information, thus helping avoid overfitting.

In a second example, the authors discuss the concept of “shortcut learning” in artificial neural networks (NN) as a type of bias implicitly integrated into AI:

> “[…] deep neural networks are often only superficially successful and fail when presented with new datasets since they learn shortcuts of the original dataset. One example is a network that learned to classify X-ray images correctly and when presented with images from a new hospital, it failed completely, since it had based its classification on a hospital-specific metal token on the scan (Zech et al. 2018).” (Fabi & Hagendorff, 2022, p.11)

In a specific environment (e.g., the same hospital), the neural network’s reliance on simple cues is effective and leads to high accuracy. The authors relate this phenomenon to how humans sometimes learn, such as students studying only to pass a test rather than for a deep understanding of the subject. In both cases, the learning strategy is adapted to the immediate environment and goals. While acknowledging the success of shortcut learning in certain contexts, the authors also notice that it can be problematic when a deeper understanding is necessary.

The failure to generalise beyond the training context is a limitation, therefore, they consider the effectiveness of this inherent bias should be evaluated based on the environment and the specific goals for which the NN was trained. Thus, despite its limitations, the authors emphasise that shortcut learning has contributed significantly to the success of image classification and should not be entirely dismissed as a useful bias in other applications.

Overall, Hagendorff and Fabi argue that the benefits of cognitive human biases can hold for machines too:

> “The proposed kinds of cognitive machine biases may, similar to human biases, be interpreted as systematic misconceptions, insensitivities to probabilities, or even errors, but in order to effectively navigate and interact with complex environments and to make accurate decisions in uncertain situations, those can become an important cornerstone. To be more precise, they may help to mitigate bias-variance dilemmas,

> avoid proneness to overfitting, simulate human decision strategies in domains where this is of importance, make models more explainable, utilize shortcuts for effective learning, etc." (Fabi & Hagendorff, 2022, p.18)

The authors claim that much like in the case of cognitive human biases, cognitive machine biases can be seen as simplifications or distortions that impact machine learning.

These biases are often thought of as flaws, but the authors propose reimagining their role: they argue that such biases could be key to successfully operating ML algorithms in complex and unpredictable environments. Ultimately, Fabi and Hagendorff support the idea that by incorporating certain types of biases into machine learning algorithms, we could improve their performance, much as heuristics improve human decision-making.

These views of AI bias suggest that instead of aiming to eliminate all forms of bias or regard them all as bad, we should recognise that some biases, when carefully selected and implemented, can be beneficial. They call to re-evaluate and acknowledge the significance of biases in AI to develop systems that are not just technically advanced but also cognitively and ethically attuned.

3.3 Untangling bias.

In the preceding sections, I have examined two distinct approaches to conceptualising bias: one that views bias as negative and another that presents a neutral perspective.

My aim in this section is to determine which of these frameworks is most appropriate for the implementation of the Bias Network Approach (BNA) and to justify this choice. However, this does not imply a plain rejection of one perspective in favour of the other. Rather, my objective is more pragmatic and modest—to adopt the framework that proves most coherent and useful for the specific application of the BNA.

Therefore, I will assume for the sake of argument, that bias is not necessarily negative. However, even working on this assumption, I will argue that we should exercise caution when employing neutral terminologies of bias within AI ethics. Given the ethical stakes involved in algorithmic decision-making, the sociotechnical nature of AI systems, and the complexity of the AI development ecosystem, I will argue that there are compelling epistemic and ethical reasons—which I have developed throughout this thesis—to frame biases as problematic (negative). This approach ensures that we remain vigilant about the systemic risks and ethical failures that biased AI systems can perpetuate, reinforcing the need for responsible and critical engagement with AI bias.

3.3.1 Avoiding technocentric and isolationist approaches to bias.

Several reasons underpin my argument. First, as outlined in Chapter 1, the technocentric bias narrative often frames bias as an isolated "fixable" problem. This perspective fosters an understanding of bias that is overly reductionist, treating it as a discrete technical issue rather than a phenomenon embedded within complex sociotechnical systems. Such an approach risks obscuring the broader ethical and societal dimensions of bias, thereby limiting our capacity to address its root causes effectively.

For example, algorithmic bias in predictive policing systems is often treated as a data problem, where the "fix" is to balance datasets. However, this overlooks systemic issues such as historical over-policing in minority communities, which are reflected in the data itself. Addressing the bias solely through data adjustments fails to confront the underlying social injustices.

The same applies to the example given by Pot et al. (2021), where they discuss the deliberate oversampling of data from underserved populations to address historical underrepresentation in medical datasets, particularly in radiology. While this practice aims to enhance fairness and improve health outcomes for these groups, the bias problem is not only in the training data.

The intention here should be to rectify existing disparities, not to just fix errors typically associated with bias. Hence, we should avoid conceptualising this as a positive, neutral, or "not all bad" bias, as this can influence our epistemic stance to understand the complexity of the ethical problem behind it.

3.3.2 Filter versus bias

Second, my objective, aligned with that of numerous scholars adopting a sociotechnical perspective, is to promote a paradigm shift in how we conceptualise ethical issues in AI. As articulated from the outset, the BNA aims to transform developers' thinking about bias. Central to this transformation is the recognition of bias as morally undesirable. This framing is crucial because it compels us to consider the multiple layers of influence—both external and internal—that operate within the AI ecosystem. These influences may constitute the origins of bias or reflect the consequences of structurally embedded biases within cultural norms, decision-making processes, and institutional practices. For instance, recruitment algorithms that inadvertently favour male candidates often do so because they are trained on historical hiring data reflecting gender biases in the workplace. Here, the issue is not merely technical but deeply rooted in societal structures that perpetuate gender inequality.

Therefore, it is imperative to exercise precision in how we refer to bias. As I mentioned above, employing language that frames bias as "not all bad" risks distorting the discourse. Similarly, I find it problematic to classify the deliberate introduction of biases into models or datasets, or the intentional selection and filtering of training data, as forms of "bias."

For example, as proposed by Fabi and Hagendorff (2022), they suggest that representational bias in social media can be introduced to benefit truth in the public discourse by fostering representational biases: "This means to change the methods for algorithmic measurement and determination of information relevance. […] –for instance, by favoring representations of rational, effortful, reflective interactions over impulsive interactions." (p.17) According to this view, if we restrict the AI's learning to only the most credible sources, we would be introducing a 'bias' towards reputable information, according to these alternative views.

Yet, it seems more accurate to call this a filter to ensure accuracy and reliability, not a bias. In such a case, "bias" operates as a safeguard against misinformation.

Labelling this selective process a bias seems to be a misnomer. It might be better to describe it as informed discernment, critical filtering, or targeted knowledge acquisition. The same can be said for the case of calling the deliberate oversampling of darker skin tones in AI training datasets a bias.

In this context, oversampling is a methodological choice designed to correct an imbalance and improve the AI's diagnostic accuracy for all skin tones, which is a justified approach. If there are any benefits for applying specific cognitive or other types of biases in AI, this should not change the notion of what we classify as a bias.

This selective process is a design choice aimed at promoting ethical content and enhancing the model's reliability. Labelling it as "bias" conflates design strategies with the unintended cognitive or algorithmic distortions typically associated with bias. The key distinction lies in whether the modification introduces a distortion or not. Filtering for accuracy is not a distortion—it enhances epistemic reliability. Bias, by contrast, involves systematic distortions that lead to epistemic, ethical, or social failures.

Hence, why I suggest that AI biases—whether intentional or unintentional—should be understood as systematic departures from justified norms that lead to unfair, inaccurate, or epistemically flawed outcomes.

Discriminatory design choices fall within this definition, whereas corrective interventions do not. This distinction is crucial for maintaining conceptual clarity in AI ethics discourse and ensuring that ethical efforts to mitigate bias are not mistakenly categorised as biases themselves.

3.3.3 A risky game of bias.

Given the significant implications that our conceptualisation of bias can have on AI development, an additional concern arises—one that remains largely unaddressed by the authors.

A key issue with Fabi and Hagendorff's proposal is their lack of engagement with the risks associated with introducing biases, as they operate under the assumption that these biases can be accurately predicted and effectively managed. However, biases—by their very nature—can produce unforeseen and complex outcomes, particularly when algorithms are deployed at scale.

Even when introduced with the best intentions, such as correcting imbalances or enhancing accuracy, biases can interact with social and technical systems in unpredictable ways, sometimes amplifying inequalities or creating new forms of discrimination. The assumption that bias can always be controlled underestimates the dynamic and evolving nature of AI systems, where feedback loops, shifting contexts, and emergent behaviours can lead to unintended and ethically problematic consequences.

For instance, recall the authors' case where shortcut learning is used as an example of implicit cognitive machine bias:

> "[…] deep neural networks are often only superficially successful and fail when presented with new datasets since they learn shortcuts of the original dataset. One example is a network that learned to classify X-ray images correctly and when presented with images from a new hospital, it failed completely, since it had based its classification on a hospital-specific metal token on the scan (Zech et al. 2018)." (Fabi & Hagendorff, 2022, p.11)

Shortcut learning occurs when AI systems identify patterns based on superficial or spurious correlations within the training data rather than learning the underlying structures relevant to the task. While such strategies may lead to seemingly successful outcomes in controlled settings, they fail catastrophically when the model encounters new or slightly altered data that lacks the same superficial cues.

In real-world applications, this could have serious consequences, particularly in high-stakes domains such as healthcare. For example, an AI system trained to diagnose medical conditions might rely on dataset-specific visual markers—such as hospital-specific artifacts—rather than genuine pathological indicators. As a result, when deployed in a different clinical environment, the system could misdiagnose conditions, potentially endangering patients due to its reliance on contextually irrelevant features.

This example underscores the broader challenge of unintended biases in AI systems—even when biases emerge unintentionally through data-driven learning processes, they can still produce significant epistemic and ethical risks. It further highlights why addressing bias requires more than just dataset curation or algorithmic tuning; it demands continuous evaluation, interpretability measures, and an awareness of how biases evolve in dynamic deployment environments.

The authors also mention a case for the explicit modelling of cognitive biases into algorithms:

> "Gadzinski and Castello (2020) aimed at combining system 1 and system 2 thinking by combining fast-and-frugal trees with ensembles of artificial neural nets that estimated Bayesian uncertainty, respectively. […] The model prediction of whether a loan was repaid was not solely dependent on exceeding a certain threshold in one variable. Instead, the prediction was leading to a certain probability of repayment when exceeding and when not reaching the threshold of certain variables. When applied in human decision- making, this procedure led to a reduction of overconfident predictions and helped humans build shortcuts while acquiring more data when necessary." (Fabi & Hagendorff, 2022, p.11-12)

While explicitly incorporating biases into AI models may, in some contexts, help streamline decision-making processes, this approach raises significant concerns—ones that the authors do not fully address.

Biases, by their very nature, are unpredictable in their long-term effects, and even when introduced deliberately, their consequences can extend far beyond their intended scope. While the structured incorporation of cognitive biases may reduce overconfidence or improve efficiency, it also risks introducing systematic errors and unintended ripple effects.

For instance, a bias introduced to optimise healthcare outcomes for a specific demographic might unintentionally disadvantage another group that was not adequately represented in the training data. This highlights the danger of assuming that the controlled introduction of bias can be inherently safe or beneficial. Even when a bias seems advantageous in a localised decision-making context, it may have broader, unaccounted-for implications when the system is deployed at scale.

Therefore, it is essential to proceed with caution. We must not underestimate the fundamental risks biases carry, particularly when they are embedded into AI systems under the assumption that they will behave predictably. To mitigate these risks, we must conceptualise bias appropriately, acknowledging its inherent unpredictability and its potential for harm. If biases are not misconceived as neutral or controllable, these dangers are less likely to be overlooked.

Thus, because biases often have far-reaching and potentially unforeseen consequences, especially when algorithms are applied at scale, even biases introduced with positive intentions can have negative ripple effects. This unpredictability of biases makes them inherently risky and something, I argue is better to classify as morally undesirable instead of potentially "good" or simply neutral. With this, I do not mean to deny, that in very specific cases, biases could be "re-purposed" to attain contextually beneficial outcomes. My claim is that we should continue to be cautious about the risk biases carry and, therefore, avoid conceptualisations that can lead developers to overlook said risk.

3.3.4 A definition of bias and its understanding in AI development.

Based on my analysis and criticisms previously presented, I will define AI bias (which considers the different categories discussed in 3.1), based on three critical properties, following Zhai and Krajcik (2023):

> "(a) deviation – bias measures the deviation between observations and ground truth (i.e., error); (b) systematic – bias refers to systematic error instead of random error; and (c) tendency – bias is a tendency to favor or against some ideas or entity over others. These three properties characterize the idea of AI bias." (Zhai & Krajcik, 2023, p.1)

Even though "over the last years, the term "bias" became synonymous with all kinds of unjust machine behavior in the fairness field." (Hagendorff & Fabi, 2023, p.3), there are specific properties that provide conceptual clarity as to what bias is. It is not just an isolated random error; it requires a systematic aspect. It is a tendency that implies favouritism or going against something but not simpliciter, this tendency presents in relation to another entity.

So, then how do we define AI bias? Undeniably it is not controversial to say that AI biases can be categorised in the three groups mentioned above. Biases come from different origins, and have different influences, on humans in the loop and the AI systems' architecture. More specifically, I here define AI bias as:

> *"An error and/or deviation based on a systematic tendency in favour of or against key entities (individuals or groups) and other elements within the AI ecosystem*."

Biases should then be understood not only as deviations based on systematic tendencies but also by virtue of their interaction and role within the AI ecosystem. This means that to avoid the isolationist approaches to bias criticised in 1.3, they need to be understood as part of a network of influences in AI developers' decision-making, with special attention to the unwanted tendencies they can cause.

Given this definition, it is important to stress that biases, particularly AI biases, are not merely incidental inaccuracies; they represent a consistent and recurring pattern of deviation from the truth. This systematic nature of bias is what differentiates it from random errors, which could happen by chance.

Biases, by contrast, are systematic suggesting a repeatable error in processing or interpretation, which, in the context of AI, can lead to the consistent and repeated misrepresentation of a problem or mistreatment of certain groups or individuals.

Regarding the aspect of bias that involves a tendency, this highlights its directional nature —it is not neutral but rather inclines in a particular (wrong) direction. In AI, this could manifest as an algorithm consistently providing favourable or unfavourable outcomes for certain demographics, thus revealing a preference or aversion that may not be justified by the data. Or a cognitive bias in professionals producing training data, that reflects biased practices that are part of the professional culture they participate in, e.g., euphemistic phrasing for diagnoses in gynaecology that can lead to distorted or inaccurate training data.

However, acknowledging these properties of bias does not provide a solution to the problem. While it is crucial to identify and understand the nature of AI biases, it is equally important to develop strategies for mitigating their undesirable effects. This involves not only technical solutions, such as algorithmic adjustments and improved data curation but also a broader societal and ethical engagement with the underlying causes of these biases.

It requires a concerted effort from different stakeholders like data scientists, ethicists, policymakers, and affected communities to ensure that AI systems operate fairly and do not perpetuate or exacerbate injustice — as portrayed in the discussion of sociotechnical views in Chapter 2.

Thus, to address AI biases effectively, we must carefully consider their systematic and tendentious nature, ensuring that our interventions are targeted and nuanced enough to address the specific ways in which AI systems deviate from different expectations of fairness, and how they relate to other biases and elements in the AI ecosystem.

I propose this general view of AI bias, because it incorporates the categories for bias previously discussed, but also expands the idea to the fact that bias happens inside or to the AI system, e.g., in the training data, in design choices by developers, in implementation and user interactions, etc. But also, outside, or indirectly related to the system, e.g., in how people do scientific research, how institutions handle data, and how systemic practices permeate problem formulation and expected solutions.

A central question remains regarding one of the main purposes of this work: how should AI developers think about bias? This is distinct from merely having a working definition of bias; the challenge lies not just in referencing a definition but in developing a framework for understanding and engaging with bias within the complex realities of AI development.

In previous chapters, I have criticised perspectives that focus solely on bias mitigation and fairness as overly narrow and insufficient (outlined in Chapter 1). In Chapter 2, I expanded on this argument by advocating for a sociotechnical approach—

one that situates AI developers and the concept of bias within a broader ethical, institutional, and social context. Furthermore, in this chapter, I have examined cases where certain biases might seemingly enhance AI system design and application. However, I maintain that there are no “good biases.” Instead, what may appear as beneficial biases are either biases that have been “re-purposed” or cases where bias is absent altogether and alternative mechanisms are at play, such as in the example of representational bias.

For the purposes of the BNA, it is crucial that developers distinguish between strategic interventions and actual biases. Even if biases are being “re-purposed” in the training process, it remains debatable whether these should still be classified as biases. What is undeniable, however, is that biases imply limitations and often reflect systemic negative tendencies. Given the multiple perspectives on bias within AI ethics, I argue that AI developers must conceptualise bias expansively—recognising its place within a complex network of interrelated biases and their origins—while retaining a fundamentally negative view of bias.

On one hand, while some biases may be deeply embedded and difficult to eliminate, developers should not dismiss them as “unfixable” and simply move on.

Biases are integral to the AI ecosystem, offering critical insights into societal injustices, institutional malpractices, or cognitive predispositions. Instead of ignoring them, developers should engage with these biases as reflective indicators of deeper systemic issues. On the other hand, some biases may be more manageable, whether through mitigation strategies or by redirecting their influence to improve AI outcomes—if, and only if, their potential impacts can be accurately predicted and controlled.

Ultimately, what is most crucial is for developers to move beyond the surface-level understanding of bias and critically engage with the contexts in which biases manifest. Recognising the role of bias within the AI development pipeline is essential for making ethical evaluations that actively counteract technocentric isolationist tendencies—where technological solutions are treated as neutral and detached from broader social implications.

While I acknowledge that some arguments conceptualise bias in AI as neutral (“not all bad”) or even beneficial in certain contexts, I maintain that, in the realm of AI ethics—particularly within the framework of the BNA—it is both epistemically and ethically prudent to treat bias as problematic (i.e., morally undesirable). This perspective aligns with a holistic understanding of AI as a sociotechnical system and fosters a critical awareness of the moral and societal implications of bias. By adopting this approach, AI developers are encouraged to engage in more responsible, reflective, and ethically informed practices.

Moreover, maintaining a clear distinction between bias and deliberate design choices is essential for effectively identifying, analysing, and addressing the ethical challenges posed by AI technologies. By drawing this distinction, we can better promote fairness, accuracy, and just outcomes in AI deployment, while avoiding conceptual conflations that might obscure the ethical stakes of AI bias.

In the next chapter, I will introduce the Bias Network Approach as a sociotechnical intervention designed to help AI developers critically reflect on bias in AI and address the challenges outlined in the first half of this thesis.

Chapter 4: A Bias Network Approach to Promote Ethical Assessments by AI Developers

In previous chapters, I discussed three main issues in the AI ethics literature. In Chapter 1, I criticised three practices about AI bias: (i) the problem of technocentrism (inclination to prioritise technical strategies in opposition to sociotechnical ones), (ii) the bias-centric view of fairness (prevailing use of bias and bias mitigation goals to define fairness), and (iii) the isolationist approach (seeing biases and mitigation strategies as individual instances in the AI developing pipeline).

In Chapter 2, I introduced sociotechnical views of AI—understanding AI systems as socio-technical systems that interact with and impact society, culture, and human behaviour in intricate ways. Amongst the limitations I examined, I noticed some gaps, particularly related to developing sociotechnical approaches that aid AI developers in their professional roles, not as an add-on or conceptual discussion, but as part of an embedded practice.

Aligned with concerns derived from this lack of embedded practice, a big part of the sociotechnical discussion in AI ethics focuses on contexts and applications that involve a high-level impact, affecting various stakeholders, particularly those most vulnerable. Healthcare, for example, is widely recognised as a high-impact area for AI ethics due to the profound implications AI technologies have on both patient care and the healthcare system.

Some of these implications stem from patient well-being. For example, when using expert systems for treatment recommendations, there are several concerns about data privacy and security of sensitive medical data, and potential inequalities in training data leading to disparities in treatment. AI can also be developed to optimise healthcare systems', by optimising resource allocation, e.g., prioritising patient needs or bed availability. These types of AI solutions have the potential to provide a great alternative to public healthcare systems that struggle with bureaucratic procedures to identify these needs.

This is the case of the "waiting list" in Chile's public healthcare system. The waiting list is an official mechanism used by the Ministry of Health in Chile and applies to all public hospitals and care facilities in the country. The waiting list comprises all referrals associated with diseases that are not covered by the Explicit Health Guarantees (EHG, abbreviated GES in Spanish), which typically includes most types of cancer and diabetes. The EHG has strict appointment timeframes and referrals with corresponding specialists. In contrast, those on the general waiting list have lengthy waiting periods to secure referrals with specialists.

Thus, the waiting list for medical and dental appointments in public hospitals in Chile represents a critical problem in the country, with significant implications for patients' well-being. As of 2017, the average waiting time for appointments exceeded 400 days, and more than 1.5 million individuals had pending referrals for appointments (Estay et al., 2017). Tragically, in 2022 alone, over 19,943 patients lost their lives while waiting for their first consultation with a specialist or a surgical intervention.[9]

As a response to these worries about putting AI ethics into practice, especially in high-impact contexts that tend to involve a multiplicity of factors influencing the development of these AI systems, I have developed the Bias Network Approach (BNA). The BNA is a transitional intervention to aid AI developers in their assessment of AI bias from a sociotechnical perspective, i.e., aiding them to transition into including broader contextual factors in their decision-making process.

Accordingly, in this chapter, I will introduce this proposal and present the main findings of a pilot case study called "the waiting list project," used to test the BNA. The pilot consisted of applying a retrospective examination of the development process for a Natural Language Processing (NLP) model used to identify key entities, such as diseases and medical procedures, in the medical and dental referrals that constitute the waiting list for appointments in public hospitals in Chile (Báez et al., 2022).

[9] As shown in the report GLOSA 06 by the Chilean Ministry of Health, https://www.minsal.cl/wp-content/uploads/2021/05/Glosa-06-III-Trimestre-2022.pdf

To introduce this approach, the chapter is divided into three sections. In section 4.1, I will introduce the Bias Network Approach, explaining its core features, and emphasising its utility as a novel visualisation and mapping tool to track elements and factors influencing bias concerns in the development process.

In section 4.2, I will introduce the pilot case study, and the methodology used for it, based on a qualitative analysis to gather insights from the developer team's experience applying the approach.

In section 4.3, I will analyse three main findings from the pilot case study: (i) the benefits of the BNA at experimental design and revision stages, (ii) the decisive role of material limitations and external decisions as bias sources, and (iii) the importance of professional biases detected with the approach. I will complement the analysis of this last finding with two concepts in the applied ethics literature: microscopic vision, drawing from Davis' work (1998) and professional deformation, based on Polyakova's analysis (2014).

4.1 The Bias Network Approach.

First, I will explain how the Bias Network Approach came to be and then characterise its core elements.

4.1.1 Background.

In the initial stages of the literature review for this thesis, as reviewed in Chapter 1, I identified isolationist tendencies to identify bias and bias mitigation. Isolationist solutions often have two limitations in practice: (i) they limit the consideration of contextual concerns during decision-making processes, contrary to sociotechnical approaches to AI systems, and (ii) they can result in a recurrent instance of risk, because if biases are seen as individual instances that need mitigation, then the same bias or a derived effect that caused a bias in earlier stages could reappear in later development stages.

Drawing from the sociotechnical literature on AI ethics, and particularly from the approach by Draude et al. (2019) discussed in section 2.4, I decided to integrate factors and elements found in these proposals to contextualise ethical analysis (including both technical and social considerations) into an approach that would allow AI developers that have limited ethical training, to engage with some of the complexities highlighted by sociotechnical views.

Hence, I wanted to see if there could be a way to introduce the network essence of sociotechnical systems as conceived in sociotechnical systems theory (STS) (see Figure 5), without neglecting or overshadowing the technical concerns rooted in the AI bias discussion, and still integrating some of the complexities of intersectionality and the interdisciplinarity demands researchers recognise as necessary elements. As a starting point, I used the illustration of a sociotechnical system by Davis et al. (2014).

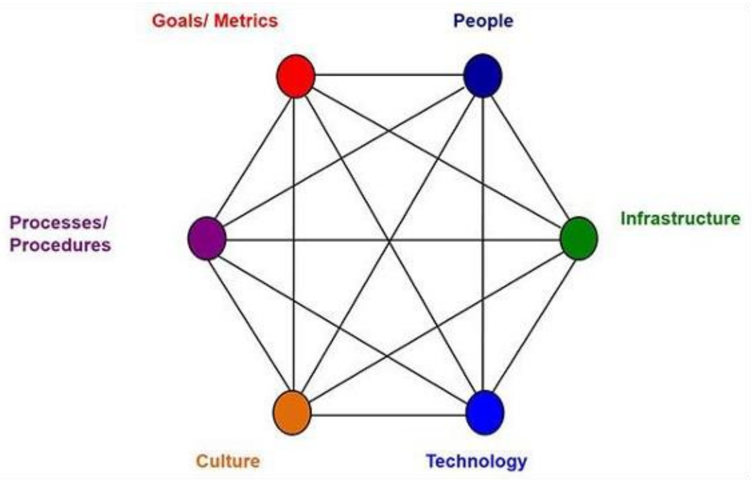


Figure 4: A sociotechnical system illustration highlighting the interrelated nature of an organisational system based on Davis et al. (2014).

Davis et al.'s (2014) illustration shows how a sociotechnical system integrates interconnected arrangements where social and technical elements interact to achieve certain goals or functions. This means, recognising that technological and social aspects are tightly intertwined, and changes in one can significantly impact the other. In a sociotechnical system, technological and social aspects are considered interdependent and inseparable elements. Changes in technology can lead to changes in social behaviour, and vice versa. With this approach, I intend to avoid viewing technology and humans in isolation and rather acknowledge their interplay and interdependence. And the same complicity applied to the interaction of humans, technology, and biases.

Accordingly, the BNA is intended to help avoid an isolationist approach by mapping biases and other relevant factors through interconnected nodes, that have direct or indirect influences. And, instead of considering organisational contexts (as depicted in the original STS Figure 5), it is applied following the pipeline structure for AI development presented in Chapter 1 (cf. Figure 2).

4.1.2 What is the Bias Network Approach (BNA)?

I have explained how the BNA was inspired by the sociotechnical illustration of a system's network of influence. But what exactly is the Bias Network Approach? First, I will clarify what this approach is not. It is not a regular protocol checklist or a committee revision. The aim of applying the BNA approach is not to assess or evaluate how well a framework was applied or to make sure that bias mitigation strategies were properly used. Instead, it is intended to be applied as an intervention to aid AI developers in integrating contextual factors in their decision-making process across the development pipeline, in this case, to identify and address AI biases.

Therefore, the Bias Network Approach is collaborative and people-based, grounded on an interactive dialogue between the AI developers and a prompter team. The latter aids the developers in creating a network map to visualise the contextual considerations relevant to dealing with bias in the AI development process and to then inform, analyse, and evaluate their decision-making processes.

To achieve this, the intervention in the case study required the following foundational elements:[10]

[10] I have called these foundational elements because this can change with other case studies, involving different developers coming from various disciplinary areas, applied in other contexts aside from healthcare, and considering that the external prompter team can also vary.

- A developer team: The AI developer team does not refer to a developer as someone involved mainly in data collection and analysis, or model testing (e.g., a data scientist outsourced or asked to do specific technical tasks). The term AI developers used here concerns researchers working on AI development projects, fulfilling roles that require technical, methodological, and procedural decision-making.

For example, an AI developer decides whichdata to use for the aim of the project, whereas the data scientist will be the onerunning analyses on the data that was given to them. In research projects, however, it is not uncommon to find AI developers fulfilling both roles. Thus,the developer team is the one in charge of designing, developing, and testing amodel or AI system to be implemented (but not necessarily being the ones implementing it).

- A group of external prompters: The network approach centres around dialogue amongst the developer team, and to guide this there is a group of external prompters. It is important to notice that this externality is not necessarily institutional, colleagues or other professionals and experts outside the developer team's institution can assume this role. The role of the prompters is to facilitate the discussion, guiding the team through relevant elements to allow the developer team to identify issues. Hence, the prompters conduct at least one semi-structured interview whilst mapping the network based on what the developer team identifies.

Follow-ups by the prompters are intended to promote more in-depth reflections about the developers' decision-making but are not intended to evaluate or intervene with the teams' insights. In other words, the prompters do not tell the team how to approach their issues but rather help them identify and map elements and factors that can help the developers visualise those issues and make better (or at least more informed) decisions.

Nevertheless, ideally, considering the type of prompting (emphasising sociotechnical considerations) the external prompters should have expertise in areas such as sociology, philosophy, anthropology, psychology, or ethics, and at least have some interdisciplinary experience to guide the discussion more effectively.

- The visualisation map tool: A final core element of the approach is the illustrative network map (Figure 6), used to visualise potential links between biases and other relevant factors. To map potential biases, the illustration process carried out by the prompters utilises the basic AI pipeline previously presented in Figure 2.

Although part of the idea of a sociotechnical view is that things are not necessarily linear, the use of the typical technical structure of the AI pipeline is to make it more accessible for AI developers who can easily recognise it as part of their existing practices.

Thus, the illustrative network is built around the AI pipeline for easier reference, but the connections do not have to follow the same temporal line. There are broad colour-coded categories to visualise biases, using the canonical categorisation introduced in section 3.1.

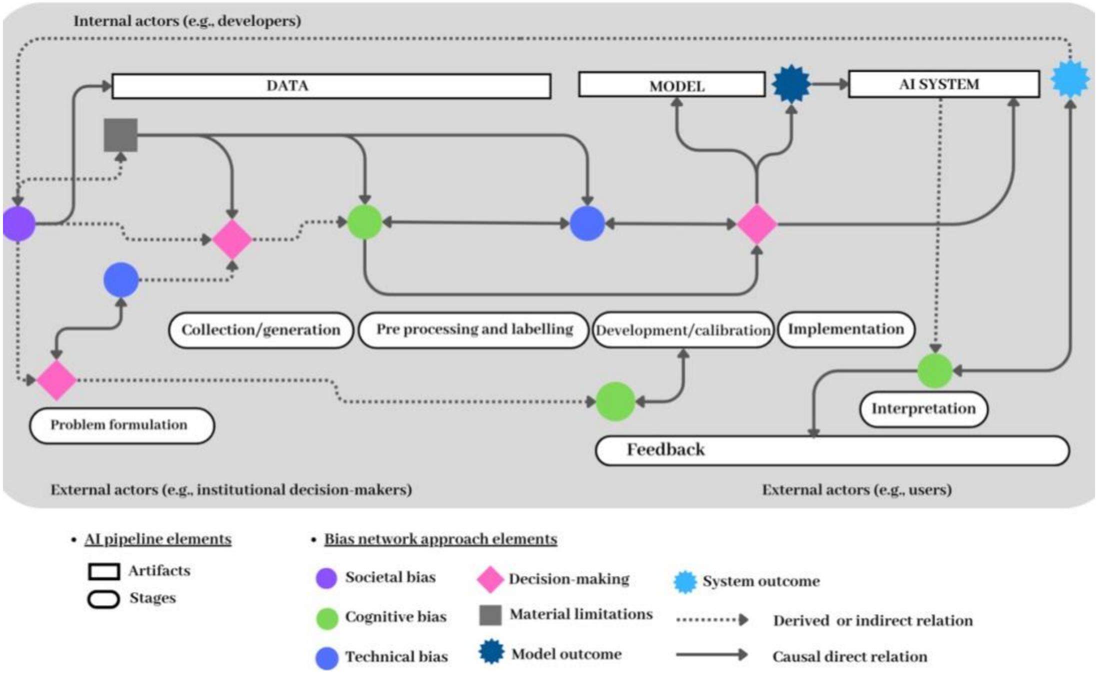


Figure 5: Illustrative network map for Bias Network Approach.

Additionally, the illustrative map includes two other elements to track influential factors that could potentially create, promote, or strengthen biases: decision-making and material limitations.

Decision-making refers to the objectives, identification of problems, assumptions, observations, principles, and general conclusions made by humans. This includes the decisions of internal actors (e.g., developers and users) and external actors (e.g., governments, data providers, and hosting institutions).

These are considered potential sources of bias, that could derive from societal or cognitive biases and affect the model's development by recreating existing biases or originating technical ones. In the case of material limitations, these can interfere with the freedom of access to data and other valuable resources (monetary and human, i.e., workforce), disrupting or limiting the development of an AI project.

These material limitations can often be attributed to external influences on the developer team, affecting the type of decision they can make or the mitigation solutions they can consider when dealing with biases.

The map also includes two different outcomes, because there are two distinct phases, and each phase culminates with a particular outcome: the model outcome for pre-deployment stages and the system outcome for after-deployment stages. Finally, there are two types of relations to link the enumerated elements in the map: direct or causal, and indirect or derived. This was introduced mainly to recognise when and where certain biases could be strongest and, therefore, harder to mitigate or deal with.

It is important to stress that this network approach does not constitute a static cartographic representation or a predetermined set of elements readily transferable across diverse contexts. Consequently, its inherent value lies in its malleability to suit the requirements of individual projects, attuned to their unique development trajectory.

Accordingly, future testing of the BNA should include projects at different development stages, as well as contexts of application, to identify further benefits and potential pitfalls.

In the following section, I will show one instance of how this Bias Network Approach can be used, by introducing the pilot case study.

4.2 Pilot case study: “The waiting list project”

To test the Bias Network Approach, I started a collaboration with colleagues[11] from the National Centre for Artificial Intelligence in Chile (CENIA), under a project called “Tackling biases: applying a Bias Network Approach to AI system development”, or as I will call it from now on, “the bias network project.”

The pilot case study selected had to have high social impact and, to be sensitive to various contextual factors through the development stages. Hence, I contacted researchers working on AI in Healthcare.

4.2.1 About the waiting list project.

The model developed by the members of the waiting list project focuses on a common task in NLP, which is named entity recognition (NER), used to automatically identify pieces of information (entities) in natural language text.

Named entities are specific entities that have names, such as people, places, organisations, dates, numbers, etc. NER tasks extract and classify these entities into predefined categories, providing structure and context to unstructured text data. In the healthcare domain, this is extremely helpful considering the abundance of unstructured electronic health records and clinical notes.

To train their model, they first needed to acquire the data. The developer team requested data from the central waiting list of 29 healthcare services in the country using Chile’s Transparency Law.[12] The positive responses to the request accounted for 23 healthcare services, leading to the acquisition of datasets spanning from 2008 to 2018. Relevant distinctions about the data distribution (that are particularly relevant for the bias network analysis performed) are:

[11] This pilot case study has been pursued in collaboration with Claudia López (second author) and Alexandra Davidoff (research assistant) and it was funded by the Centro Nacional de Inteligencia Artificial CENIA, FB210017, BASAL, ANID. The analysis of the pilot case study is published in an article comprising primarily Chapters 4 and 5 of the thesis (Arriagada-Bruneau et al., 2025)

[12] Ministerio Secretaría General de la Presidencia. 2008. Bill 20.285. https://www.leychile.cl/Navegar?idNorma=276363&idParte=

The dataset encompassed a total of 5,157,902 referrals, originating from 40 distinct medical specialties and 11 dental specialties, in line with the country's regulatory framework to classify specialties on the waiting list.[13]

The division between medical and dental referrals exhibited a distribution of 88% representativeness from medical records and only 12% for dental ones. The dataset ended up containing 994,946 distinct diagnostic terms. A subset of diagnostic terms, totalling 107,235 unique candidates, was selected by the team for annotation based on a criterion of using only those diagnostics exceeding 100 characters.

Diagnostic terms with anomalies or lacking additional information were excluded after scrutiny by a managerial authority, ensuring compliance with set criteria and safeguarding personal information on top of the anonymised nature of the original data.

Afterwards, the team started the annotation process. This was conducted by a team of five annotators, including medical students, a medical doctor, and a dentist. The selection of annotators with diverse backgrounds was intentional and considered important to assess the variety of medical specialties in the dataset. While clinically trained annotators were shown to excel in annotating clinical text, non-medically trained annotators were also found to achieve significant agreement in semantic annotation.

The annotation procedure consisted of two stages. In the initial stage, guidelines were formulated through an in-depth analysis of existing comparable guidelines and evaluated during the annotation of a subset of referrals, leading to the creation of a reference set. The second stage involved annotating 50 referrals by the medical students over three weeks, incorporating iterative training and guideline enhancement, culminating in the establishment of accepted guidelines. A manual pre-annotation phase was introduced for basic entities, carried out by medical students, allowing senior annotators to focus on more complex entities and relationships requiring advanced clinical knowledge.

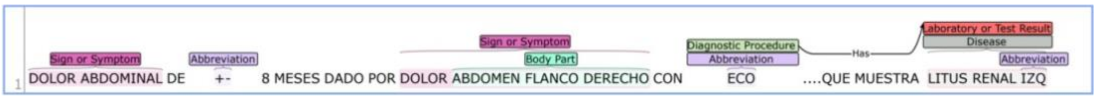
[13] Technical guidance for the registry of the Chilean waiting list, provided by the Ministry of Health. https://www.minsal.cl/wp-content/uploads/2016/03/Norma-Tecnica-118.pdf

To consolidate annotations, they were reviewed by a panel of four researchers, including a senior annotator, dentist, guideline manager, and principal investigator. Through collaborative analysis and discussion, consensus was achieved for each annotation, integrating them into the established ground truth.

The ground truth serves as a benchmark or standard to assess the correctness or reliability of measurements, models, algorithms, or annotations. In the context of the waiting list project, "ground truth" refers to the set of annotations or diagnoses that have been manually reviewed, verified, and agreed upon by a panel of experts. These annotations are considered accurate and trustworthy, providing a solid foundation against which automated or algorithmic annotations can be evaluated for accuracy and alignment.

The final ground truth encompassed 2,067 dental and 2,933 medical annotated referrals, indicating an oversampling of dental referrals in the ground truth. Specifically, dental referrals accounted for approximately 41% of the ground truth, even though they constituted only 12% of the referrals within the complete dataset (Báez et al., 2022). The team ended up formulating a comprehensive codebook to classify seven distinct entity categories, namely: disease, body part, medication, family member, abbreviation, procedure, and finding.

A medical example of this annotation is given by the authors, they used the BRAT Rapid Annotation Tool to make these.



DOLOR ABDOMINAL DE +- 8 MESES DADO POR DOLOR ABDOMEN FLANCO DERECHO CON ECO ....QUE MUESTRA LITUS RENAL IZQ

Figure 6: Example of a medical referral using BRAT.

In Figure 7, the full sentence translates to: "Abdominal pain + - 8 months due to abdominal pain on the right side with an echo that shows kidney stones on the left side". "Abdominal pain" is categorised as a sign or symptom "+ -" is an abbreviation, "pain" is categorised as a sign or symptom, "abdominal pain on the right side" is a body part, "an echo" is recognised as an abbreviation and a diagnostic procedure, which is then connected to a laboratory or test result and a corresponding disease “kidney stones on the left side.”

Then, to build the model, the waiting list project team used a Multiple Single-entity (MSEN) approach for entity recognition in the referrals. They applied it to find named entities within text that are nested inside each other. For this, they trained individual models, each focusing on a specific type of entity. To make each of these models, the developers of the waiting list project used an approach called LSTM-CRF.

The approach references Long Short-Term Memory (LSTM) networks and Conditional Random Fields (CRF). On the one hand, LSTM helps the neural network model remember entities that are far apart, as well as having a sort of “gatekeeper” to decide which information is important to remember and which can be forgotten. CRF, on the other hand, helps the model figure out the best labels for each word or entity in relation to nearby ones. CRF looks at how words in a sequence are connected, it is good at finding these patterns and connections. This combined LSTM-CRF architecture takes elements from both LSTM networks and CRF models for sequence labelling tasks.

In simple words, think of this as having different experts, each specialised in finding a certain kind of thing (entity or category) in a sentence. These experts work together to find all the things mentioned in the sentence. This way saves time and makes finding nested entities easier. To achieve this, the waiting list team turned the input sentences into useful forms for the experts.

By creating different ways of representing words, they are broken down into smaller parts. At the end of this process, the team used an algorithm to figure out the most likely way to label each word in the sentence based on the patterns learned. This helps find the named entities and the relationships between them, following a specific labelling format. The final MSEN model incorporates a sequence labelling model, and clinical word embeddings concatenated with character and Flair embeddings (known for capturing contextual information from text, allowing NLP models to understand the meaning of words and sentences in relation to their surrounding words).

The model gave outcomes with a resulting F1-score of 80.27, significantly surpassing the baseline model (a layered neural model). A score of 80.27 means that the model is performing well in terms of both precision (how many of the predicted positive cases were correct) and recall (the proportion of actual positives that were correctly predicted by the model), striking a good balance between these two metrics.

It is important to note that, as this work was being written, the model of the waiting list project has not yet been widely deployed in the healthcare system. The research team is currently collaborating with the Ministry of Health to further refine the model and AI systems that will underpin clinical management decision-making. As they continue to work towards the integration of their findings into practical healthcare settings, the potential impact of their project in optimising clinical processes and resource allocation remains highly promising. Precisely because of this, they have agreed to test their model with the BNA to reflect on areas for improvement and gain awareness for future evaluations of the model itself and its implementation.

4.2.2 The methodology for the case study.

In this pilot case study, a qualitative research approach was used, aiming to explore the subjective experiences and interpretations of participants concerning their utilisation of the BNA within a retrospective evaluation of their decision-making processes.[14]

The primary objective was to acquire a fundamental element of qualitative research: "to provide in-depth insights and understanding of real-world problems." (Moser & Korstjens, 2017, p.271) In this case, this included both the ethical concerns coming up from developing AI solutions in the healthcare domain, and the specific problems developers faced during their project.

[14] For this study informed consent requirements were fulfilled for recording the interviews at each intervention, as well as informing the developer team about the purpose of their participation, the type of analysis being performed, and the use of the findings, i.e., academic publications. No sensitive information was used or collected. The project underwent an ethical committee revision at the Pontifical Catholic University of Chile, and it was approved to run the pilot and further case studies within the same project, by the Scientific and Ethical Committee for Humanities and Social Sciences, ID of the approval 230810003.

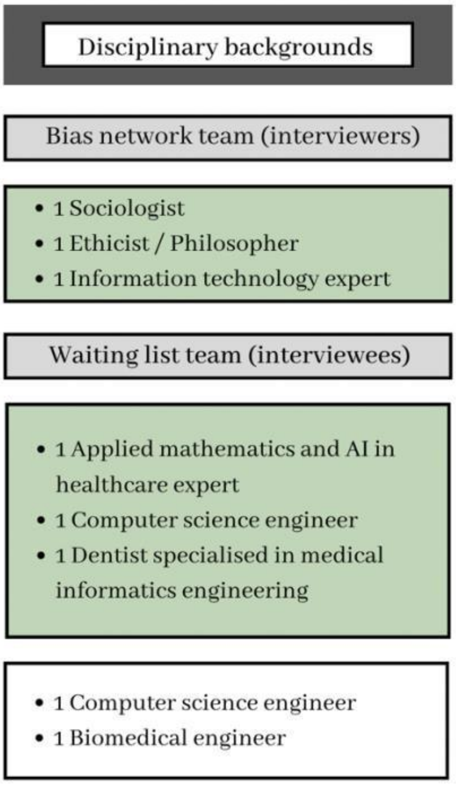

Disciplinary backgrounds
Bias network team (interviewers)
• 1 Sociologist
• 1 Ethicist / Philosopher
• 1 Information technology expert
Waiting list team (interviewees)
• 1 Applied mathematics and AI in healthcare expert
• 1 Computer science engineer
• 1 Dentist specialised in medical informatics engineering
• 1 Computer science engineer
• 1 Biomedical engineer

Figure 7: Disciplinary backgrounds of participants in the pilot case study.

Participants: For the pilot, there were 3 interviewers from the “bias network project” and 3 interviewees from the “waiting list project” (out of 5 researchers). I highlight the disciplinary background of each participant in Figure 8, as this is central to the main findings discussed later.

Sampling the case study: For the pilot, the waiting list project was sampled based on the following criteria:

(i) Have a small developer team. This was more likely to include the developer definition mentioned above, i.e., developers that make key methodological decisions (something that is less common in heavily outsourced or bigger teams).

(ii) Work with a team that already has a model developed. Although I expect the BNA to be useful during various stages of a development project, since this was an exploratory pilot, using a retrospective analysis was a more manageable approach to get some initial perspectives and insights to later test the approach on different projects at various development stages.

(iii) The model will be implemented in a high social impact context. It was essential to use sociotechnical contextualisation to identify ethical distinctions about biases, so recruiting a project that had some inherent ethical dimensions because of the nature of the data, or the context of application was key. In this case, it was healthcare.

Semi-structured interviews: The pilot involved performing semi-structured interviews as part of the intervention, as they provide a flexible yet systematic framework for analysis. In this case, the interview lasted two hours and included open-ended questions and prompts that encouraged participants to share their views and narratives behind the decisions they made throughout the development of their model and how they connected these to identified biases and other bias-related sources. Interview sessions were video recorded with participant consent, and then only the audio files were stored for data analysis. These recordings were transcribed verbatim, capturing the nuances of participants' responses. Two audio transcription programs were used, oTranscribe and GoTranscribe.

To carry out the interviews, both the ethicist (myself) and the information technology expert (Claudia López, co-author) prompted participants using the stages in the AI pipeline depicted in Figure 1, and the bias identification and mitigation mapping from the literature reviews mentioned before (see Figures 2 and 3). The prompting involved asking the "waiting list team" to discuss the decisions they made in each development stage and asking for further information regarding their reasoning behind it. For example, from the problem formulation stage, a prompt involved showing the participants a list of the most common biases identified for that stage and then following up with procedural questions such as: "How did you formulate the problem, and how does it relate to data collection decisions you made?", "To what extent did you consider any of the biases that we are showing you here? (Or others that might come to mind)", and "Did any of your initial considerations change after data acquisition?"

To avoid potential biases in this process I incorporated a list of only the most common biases in the literature, rather than focusing on biases I (as one of the other prompters) might have identified when learning about the waiting list project ourselves. Follow- ups consisted mainly about procedures and decisions that were initially brought up by the research team, allowing participants to elaborate on their thoughts and perspectives.

Partly, this was to avoid biasing the developer team to make the ethical analysis through the lens of my own experience and narrative about biases. Nevertheless, if they asked for clarification or if I could expand on definitions for certain concepts and ideas, they were unfamiliar with, further information was given, e.g. if they wanted to know more about gender biases, or if there was a name for a particular experience they had or they recognise as an instance of injustice caused by bias.

Data analysis: For this, I performed a thematic analysis with the help of Alexandra Davidoff, the sociologist. We systematically grouped and organised related biases and elements, such as material conditions and decisions made by the developer team, that emerged from the interview. First, we familiarised ourselves with the data from the transcripts and personal notes from the interview process. Afterwards, we performed a specific coding analysis for the categories of bias sources identified by the participants. By utilising open coding techniques, a total of 46 codes were created (see Annex, Table 2), each characterising a specific bias source. These codes were further classified into thematic groups, facilitating a systematic examination of the data, and these thematic groups were later validated by the information technology expert in the group.

Throughout the analysis process, each bias source was carefully examined, and explicit connections with other sources were identified whenever the developers of the waiting list project mentioned interrelationships between them. For each theme, we selected illustrative examples to showcase what participants said or did that led to the identification of that theme (which were later validated by the other prompter, the information technology expert).

These examples provide evidence of the existence of the theme in the data, which will be presented in the upcoming findings section, by using the illustrative map shown in Figure 6 above.

Intersectionality: As discussed in Chapter 2, intersectionality is a core element grounding the theoretical framework for the BNA. Overlapping social identities and systemic inequalities interact to create unique experiences of discrimination, privilege, and oppression. Consequently, intersectionality considers the broader social, historical, and institutional contexts in which identities and systems of power operate.

In the pilot case study one clear intersectional discussion arose because of the notable challenges related to biases embedded within gynaecological data, requiring the developers to address multiple intersecting social, systemic, and professional factors.

This stressed the importance of incorporating intersectionality, as part of the theoretical framework for grounding the BNA, as this can lead the developer team to identify how overlapping and interdependent systems of discrimination and inequality influenced the dataset, the model' performance, and the broader healthcare context.

During the application of the BNA, biases in gynaecological data were traced back to both professional and systemic biases present at the intersection of healthcare practices, gender norms, and resource inequities:

Gendered Practices in Medical Documentation: Gynaecological diagnoses were frequently handled by midwives or nurses, but male physicians often used vague or informal terminology in their documentation. This inconsistency in terminology and practice reflects entrenched professional hierarchies and the gendered undervaluation of women's health, leading to disparities in data annotation and a lack of standardization in medical records.

Systemic Inequities in Healthcare: Primary care establishments, funded at the municipal rather than central government level, exhibited disparities in resources and data quality. This systemic inequality compounded the challenges of training the model with representative and high-quality gynaecological data.

4.3 Analysing the main findings of the case study.

Here, I will discuss three main findings from the case study: (i) the benefits of the Bias Network Approach at the experimental design and revision stages, (ii), the decisive role of material limitations and external decisions as bias sources, (iii) and the importance of professional biases detected with the approach. I will illustrate and analyse the connection made by the developers as well as highlight the benefits they identified after applying the BNA.

These findings, given the nature of the qualitative research approach of the case study, are not expected to be generalisable results, however, some of the general insights derived from these results show promise for applying the BNA as a transitional intervention.

The emphasis is put on providing a perspective about the concerns and benefits of using this network approach brought up by the developer team. Based on these findings, nonetheless, I intend to design further case studies with other projects to identify tendencies regarding the general benefits of applying the BNA, its limitations, improvements it can offer to the developers, and the individual professional influence the BNA can have on developers' awareness of their ethical decision-making process.

Here, I will discuss three main findings from the case study: (i) the benefits of the BNA at the design and revision stages, (ii), the decisive role of material limitations and external decisions as bias sources, and (iii) the importance of professional biases detected with the approach. In discussing the findings, I will identify participants as Developer A, Developer B, and Developer C, mainly to give the same importance to all their interventions (by not distinguishing a principal investigator, for example). The direct quotations by each developer have been slightly adapted in the Spanish-English translation, as the BNA case study intervention was conducted in Spanish.

4.3.1 Insights by the developer team: visualisation, collective discussion, and transparency.

Here I will discuss findings and insights related to main findings (i) and (iii). During the coding process for data analysis, it was possible to observe that the thematic groups connected different sources of multiple types of bias, linking decisions, influences, and limitations throughout the pipeline, i.e., there was an interconnection across stages. It was possible to map different issues as depicted in Figure 9 below.

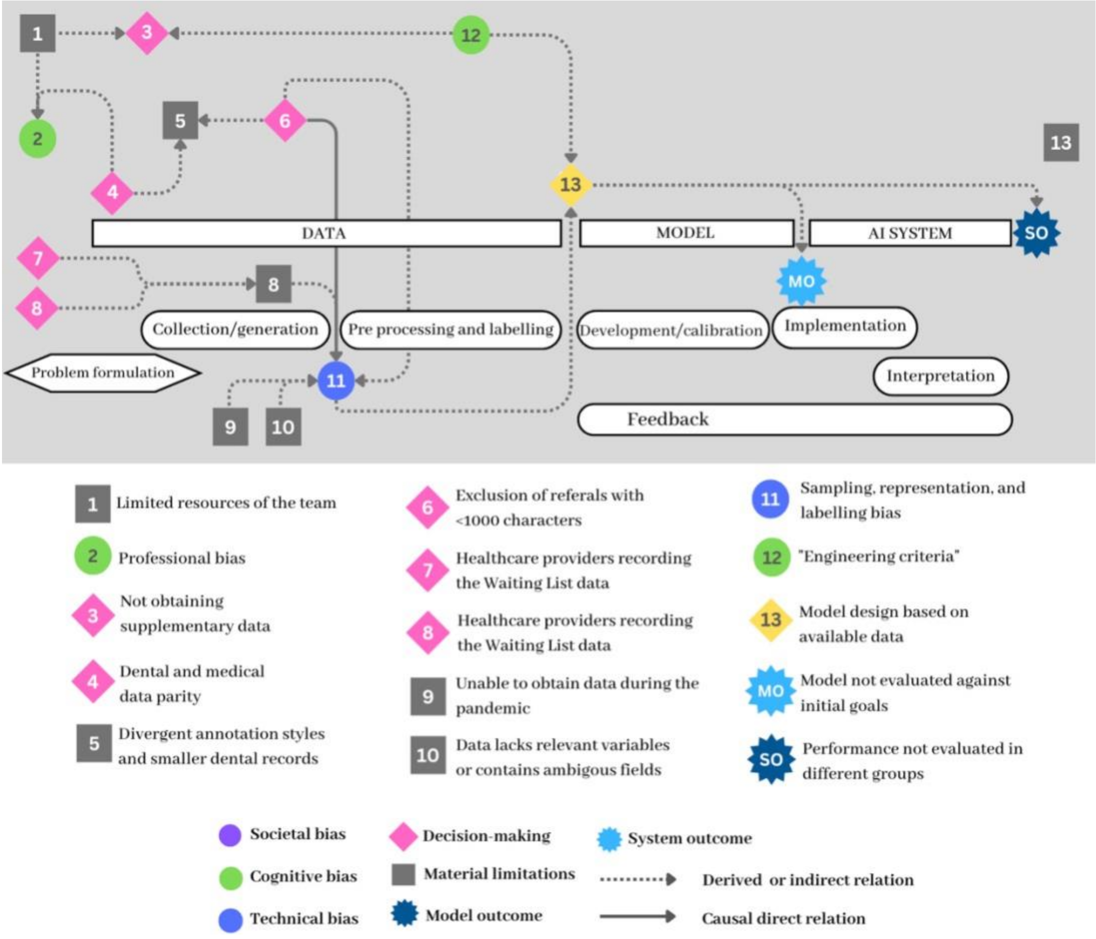

1
3
12
2
5
6
13
4
13
DATA
MODEL
AI SYSTEM
SO
7
8
MO
8
Collection/generation
Pre processing and labelling
Development/calibration
Implementation
Problem formulation
11
Interpretation
9
10
Feedback
1 Limited resources of the team
2 Professional bias
3 Not obtaining supplementary data
4 Dental and medical data parity
5 Divergent annotation styles and smaller dental records
6 Exclusion of referals with <1000 characters
7 Healthcare providers recording the Waiting List data
8 Healthcare providers recording the Waiting List data
9 Unable to obtain data during the pandemic
10 Data lacks relevant variables or contains ambigous fields
11 Sampling, representation, and labelling bias
12 "Engineering criteria"
13 Model design based on available data
MO Model not evaluated against initial goals
SO Performance not evaluated in different groups
Societal bias
Cognitive bias
Technical bias
Decision-making
Material limitations
Model outcome
System outcome
Derived or indirect relation
Causal direct relation

Figure 8: Network map showing the contributing factors influencing model design based on available data as identified by the developer.

For example, numbers 1, 8, and 9 were important limitations experienced by the developer team:

> Developer B: “We didn’t have data, and when I say we didn't, I mean it. In the beginning, we did not have an agreement with the Ministry of Health; we were in the middle of the pandemic, so nothing was available. So, what we did was ask for the data through the transparency law, but even then, the healthcare institutions were not forced to give us any data. There is an exemption that states that if it is too time-consuming, they can opt to not provide any information. So, we had to beg, like, "Please, we want to help, so send us anything you have or anything you can.”

In this scenario, the waiting list team was just relieved to have data to work with. Accordingly, as we were prompting them with potential biases in that data acquisition stage during the interview, one developer noticed that:

> Developer C: “They [the hospitals, or healthcare centres] sent whatever they had. We had no clarity about the timeframe for the data. [...] But yeah, there were important differences. Like if they had access to digital data, it was easier or better structured, the thing was that there was no uniformity in how they recorded them, some centres keep manual databases as a parallel tracking system, so we could not really know.”

This challenging context for obtaining data and the type and quality of data they were able to obtain created further issues. The developers noticed that there was a bias in how they decided to put a threshold on the selected training data:

> Developer B: “There was this criterion based on how people wrote a diagnosis. They all tend to be short, so we decided that we would use only the ones with 100 characters, we selected referrals like that, and those were the ones we finally annotated.”

Recognising that the quality and style of records significantly vary across different specialties and healthcare providers, the decision to only select records of a certain length played a pivotal role in shaping their final model. By excluding shorter records (which are more common in dental entries), both the labelling process and the functionality of the algorithm were influenced.

Consequently, the data selection criteria could introduce a bias that may have impacted the model’s performance and the quality of its outputs, particularly in contexts where shorter records are more prevalent. The team identified it as a potential bias, and emphasised the interconnectedness of their own decisions with external limitations:

> Developer A: “We decided to have dental and medical data parity, but we could be introducing a bias by overrepresenting dental data. Because we do have more medical data, there are also more of those in the system, so I don’t know, this is a bias that comes to mind just now. In our defence, we really wanted to include dental data

> because there is a significant number of people on the waiting list associated with dental needs, it is like relative importance […] in the end, I also think there was an idea of using the limited human resources we had, having a dentist, we kind of said we need to use that."

Developer C complements:

> "Yeah, actually, the difference between medical and dental data was big. There were 42 words on average used in medical diagnoses and only 28 in dental ones. I think we just put it all in a table, but we did not make any conscious decisions about it, it was not part of "building the model" it was just data sampling, but there was a bias in that. And it goes back to all those other things we discussed. Actually, if you think about it, we are just applying what we were able to gather, without really thinking about how the data gathered by other professionals is affecting how we think about our model."

This was mapped in the interaction of numbers 4, 5, 6, 11, and 13 in Figure 9.

Another pronounced bias noticed by the developers was related to the notably inferior performance of the model when confronted with gynaecological data. As illustrated in Figure 9 where numeral 2 symbolises a professional bias exerting an influence on numeral 5, the divergent annotation methodologies correlate with a concern specifically observed within the domain of gynaecology:

> Developer B: "Our model, with our current data, performs great on average, but the worst performance is for gynaecological data. Just 1/3 of the time it gets it right [referring to the performance score]."

To elucidate potential shortcomings, the same developer shares their perceptions regarding factors that might affect the model's efficacy in handling gynaecological data. Revising their documentation and stored data, they discerned a connection between this issue and alterations observed in the recorded suspected diagnoses within the domain of gynaecology:

> Developer B: "Most people in other fields [medical specialties] write a diagnosis, and it is normally reported by the physician. So, in gastroenterology, for example, it directly identifies the need for an endoscopy. But, in gynaecology is not like that. Actually, I consulted some physicians, and most of the diagnoses are written by midwives or nurses, so they do not actually write a diagnosis. And when they do, male doctors tend to underestimate symptoms or write in non-medical terms. For example, refer to "the region" instead of the clinical term. It is so freakish that this can affect the process [referring to the developing process]."

In line with this last comment, a recurrent reflection emphasised by the participants was thinking about the limitations and decisions performed by them as developers (thematic group 5 in the book of codes) and the fact that the BNA allowed them to think about it collectively:

> Developer A: "The best thing is the fact that someone external is asking us these questions and guiding us, and that it is not heavily structured, because problems were flowing and we could have a conversation, so we were able to naturally go to the things we thought were more important."

Furthermore, the same developer noticed that there were potential benefits to implementing this approach like they did (retrospectively), but also from the beginning, applying it before starting their data collection:

> Developer A: "This could be implemented at two levels, in my opinion [pointing to the illustrative maps resulting from the bias network analysis, e.g., Figure 5]. One is the experimental design phase because asking all these questions and doing this sort of group analysis with the team makes the research better. The other one is to like make things transparent, potential biases that my results have (that I perhaps cannot change), but that are worth publishing."

Developers also noticed how the visualisation aspect can help them in managing the networking process of building the bias network:

> Developer B: "It is pretty overwhelming all of this [referring to thinking about biases], because the more we think about it the more biases we identify and connect. So, having a way to map them in a process makes one think straighter and not get lost. For example, it just came to mind the case of primary care. These establishments are funded by municipalities, not by the central government. So, there is an unavoidable inequality in the distribution of resources, the municipalities with more money have better healthcare and better systems to manage the data. And there you go, another bias!"

> Developer A: "I mean, [identifying biases] is like one part of learning, because when one really needs to be aware of something or be conscious about a problem, having different modes of analysing that information is what helps. So, we had the interview and prompts, we got some insight from that, and then we also had visual support with the network maps, so having both things is what makes it better, that is more important… it is not how information is being transferred but the fact that all these things are done simultaneously."

Analysing the developers' answers, we identified a tendency to realise diverging interests among different stakeholders as another societal factor that impacts their project, particularly regarding data-sharing decisions made by institutional external actors. Beyond individual healthcare providers' deliberate choices, the overall structure of the healthcare system, also constitutes a significant determinant in the emergence of bias, as shown in Figure 9.

The referrals obtained by the developer team were not intended to be used as training data for an AI model but rather to document operational processes and meet institutional objectives.

As a result, the data might not encompass all the fields and variables that the research team considered relevant. Additionally, external actors can limit or manipulate the data provided. For example, some healthcare providers may manually annotate a portion of their waiting list on physical paper, excluding it from official records to project an illusion of enhanced efficiency.

Another important finding for the developer team was recognising how the significant differences between the quantity and quality of data provided by different medical specialties presented unique challenges. This variation may stem from the unique cultures of each medical field and their respective number of medical appointments, resulting in disparities in data representation across specialties.

The field of gynaecology, for instance, raises important questions, as cultural conceptions surrounding women's health might impact the precision and vocabulary used in records. The involvement of professional midwives in filling out gynaecological records further adds complexity to the use of specialised medical terminology.

Finally, amongst the benefits they commented on was the fact of having a visualisation of these influences and the organic connection that came up in their analysis. They realised that talking about their project within this sociotechnical approach, made them recognise how external societal factors were influencing the way they approached model building and how they understood the relevance given to their technical decisions —often highlighting the things they did not consider or never thought could be a source of bias.

Likewise, they expressed affinity to having a conversation, when asked what they thought about the BNA, Developer A stressed the benefits of thinking collectively as a team and in a network fashion:

> Developer A: "There are a lot of benefits to implementing it [the approach] at the beginning and the end, because I can check problems and maybe change the course of action. Even retrospectively, it helps you think about how to model stuff, like presenting it in a way that makes you think about these issues throughout. The thing that really helps and nourishes the process is that talk with the team members is the guided conversation because is something that could really help if done even before data collection."

These insights brought up an aspect that I did not consider before the pilot case study. Part of the aim of using the BNA was to create an interdisciplinary interaction to make the process more accessible for AI developers who do not have much ethical training. I was aware that this is quite demanding and renders the approach very dependent on having resources (material and human) to have external interviewers implement it.

However, considering the type of interaction in which the team engaged naturally using the prompts, I can see the potential for the network approach to have certain structured elements that could make it applicable to a wider variety of projects.

The fact that direct and indirect connections were organically raised by the developers, gives us reason to think that —as developer A noticed— the benefit of the network approach is having multiple inputs for the analysis, i.e. external experts give prompts, but they also have a visualisation and discussion component. These elements could be broadly systematised to be applicable in different contexts, particularly limited ones in terms of the availability of external experts or a lack of institutional support.

In that same spirit of a collective discussion, developer A said that the fact the BNA was shown as a network helped because it is a better and more natural way to understand a problem:

> Developer A: "Well, it feels more organic, because that is how you should analyse a problem. If you analyse things part by part, it is easier, but you overlook the continuum, the interaction of the characteristics you are trying to systematize. Different aspects are not independent, so I think it helps, because it helps you

> generalize the analysis, so you do not have a discreet analysis of all the individual parts, but instead of the relation amongst those parts."

This was particularly relevant, as this is part of the fundamental objective of a sociotechnical approach like the Bias Network Approach, which reflects on context and avoids isolationism.

Mapping connections and links amongst relevant elements seems to be an approachable way to promote context awareness in a non- overwhelming way because the context is not tracked by a set of assumptions or expectations but by the worries and concerns identified by the thought process of the developer team itself.

Finally, a benefit that I did not foresee or expect, but that was brought up by the developer team, was the fact that the BNA could be very useful for them to track and justify the ethical challenges or the more general ethical discussion of their work for article publication or conference calls:

> Developer A: "I am not an expert in the area [referring to ethics] so for me I like simple things, something that has utility, so that I can add it to my paper and if it also helps me make my research better, that's good too. I can see how this can help you visualise and make decisions transparent and socialise deficiencies in my project or experiment. So, it is a benefit for us [developers] but also for socialising or explaining our process to the community. In the end, this helps because making transparent choices about methods, data, and other decisions helps to evaluate if something is good or bad like if the process is ok or not, you know what I mean? So, this can help make transparent the reflection of the developer team, the criteria we followed."

This made me consider how the BNA cannot only be presented as an intervention for internal purposes (better research evaluation or ethical assessments) but also offer explainability and transparency elements for developers to socialise their decision-making process, particularly ethical decisions, as part of their professional practice — something sociotechnical approaches promote.

4.3.2 Let's pay more attention to professional biases.

One thing that caught my attention when we were analysing the findings of the case study was the presence of professional bias. In the revised literature examined in previous chapters, professional biases were not part of the prominent discussion about

AI bias. There is a consistent lack of explicit and implicit mention of professional bias in the core AI literature, as this is often presented as a separate discussion.

Accordingly, this was an important finding, as it provided a different dimension of discussion for AI bias that came from the developers' reflective process.

Professional bias emerged as a recurring type of bias evident in the decision-making processes of the AI developer team —it could be attributed to a type of cognitive bias, but professional bias has unique elements that warrant a distinction. The developers acknowledged a tendency to prioritise technical criteria aligned with their professional background, primarily in the field of engineering. Moreover, they recognised that certain aspects influencing their decision-making were not considered in their thought process for important methodological decisions.

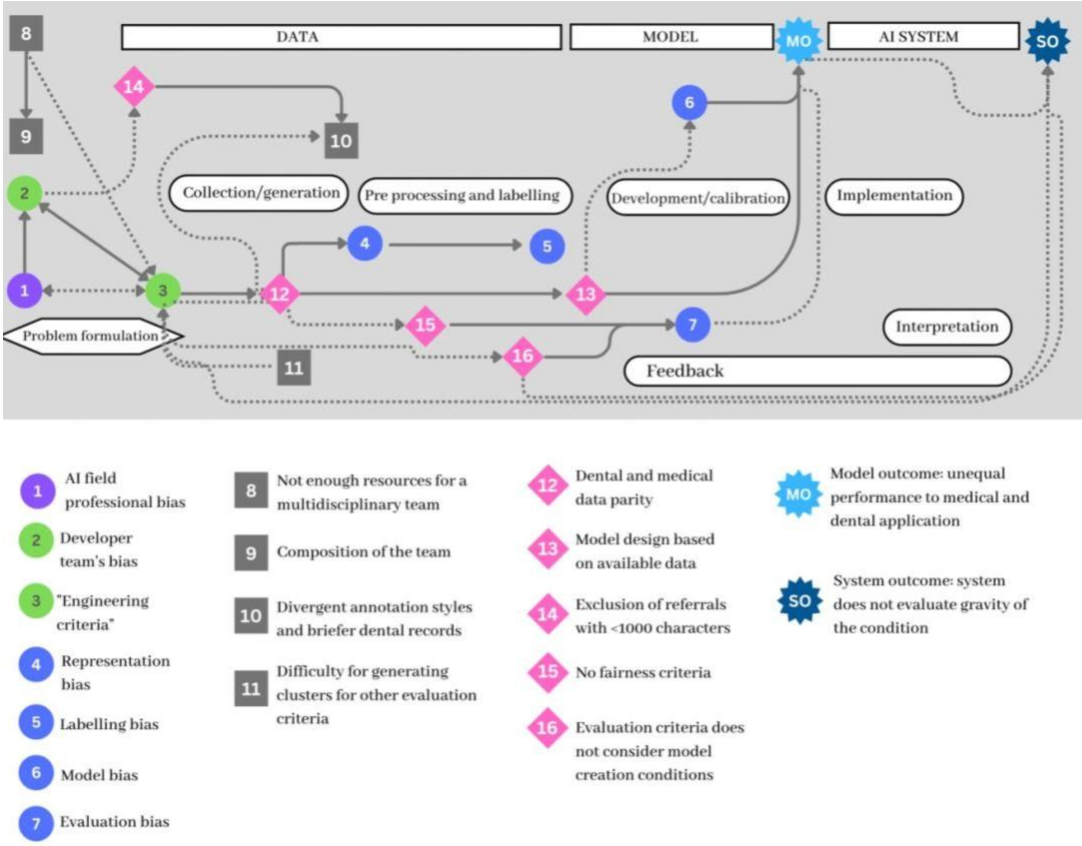

DATA
MODEL
MO
AI SYSTEM
SO
8
9
14
10
6
2
Collection/generation
Pre processing and labelling
Development/calibration
Implementation
4
5
1
3
12
13
15
7
Interpretation
Problem formulation
11
16
Feedback
1 AI field professional bias
2 Developer team's bias
3 "Engineering criteria"
4 Representation bias
5 Labelling bias
6 Model bias
7 Evaluation bias
8 Not enough resources for a multidisciplinary team
9 Composition of the team
10 Divergent annotation styles and briefer dental records
11 Difficulty for generating clusters for other evaluation criteria
12 Dental and medical data parity
13 Model design based on available data
14 Exclusion of referrals with <1000 characters
15 No fairness criteria
16 Evaluation criteria does not consider model creation conditions
MO Model outcome: unequal performance to medical and dental application
SO System outcome: system does not evaluate gravity of the condition

Figure 9: Network map showing the influence of professional bias as identified by the developers.

During interactions with the developer team, it became evident that their technically driven decisions were not intentionally disregarding ethical considerations. Instead, they were making choices based on their available expertise, which enabled them to make crucial decisions, but unavoidably overlook crucial sociotechnical factors.

The main issue was that they were unaware of how their technical decisions were interconnected with other sociotechnical aspects that could affect the AI model they were developing. These interconnections were identified by the developers, as shown in Figure 10 above.

As discussed in previous chapters, in the sociotechnical AI ethics literature, several researchers have highlighted a fundamental critique against technocentrism and technological solutionism. However, little attention has been given to understanding the origin of this technocentric tendency from a professional bias perspective, specifically examining the factors that drive AI developers to adopt those approaches. As developer C noticed, this bias can be associated with a lack of ability to look outside their technical focus:

> Developer C: "We were so focused on the technical side that we did not give ourselves any time to think about things around us or look at them from a distance, from the outside. That is why we did not have this type of discussion before."

This phenomenon has been discussed in applied ethics, particularly engineering ethics. In "Thinking Like an Engineer" Davis (1998) introduces the concept of "microscopic vision" referring to a narrow focus and detailed problem-solving approach characteristic of engineers. Davis argues that engineers tend to adopt a microscopic perspective when tackling technical challenges, focusing on specific details and technical aspects to arrive at efficient and effective solutions.

Microscopic vision is a manifestation of the engineering mindset, which prioritises technical rationality and problem-solving based on scientific principles and technical expertise. Engineers are trained to break down complex problems into smaller, more manageable components, analysing each part in isolation to find optimised solutions. This approach is highly valued in engineering practice as it leads to effective problem-solving and innovative designs. Developer B noticed this and called it their "engineering criteria:"

> Developer B: "I think there is a very strong bias present here, the fact that we all come from an engineering-related background. I mean the only thing we truly cared about was getting the best performance, the F1 score. So, we looked at the literature to see models and metrics that would help us do just that, our "engineering criteria." But never really stopped to think how that "best metric" is going to help the lady waiting for medical care to fix her eyesight?"

Davis (1998) raises ethical concerns regarding this microscopic vision. He argues that engineers' intense focus on technical aspects and efficiency may cause them to overlook or downplay broader ethical considerations and social implications. By narrowly concentrating on technical solutions, engineers might fail to recognise the ethical dilemmas inherent in their work or ignore the potential impact of their designs on society, the environment, or vulnerable populations.

In this sense, this vision involves a selective focus on specific information deemed more relevant or useful in a given situation, while disregarding less pertinent details. It can be likened to looking through a microscope, where one gains the ability to observe intricate details that would otherwise be imperceptible to the naked eye —a metaphorical mental process.

Davis also notices that it is essential to differentiate the metaphor of microscopic vision from near-sightedness or myopia, which refers to a visual impairment that affects distant vision while maintaining clear close-up vision. In contrast, microscopic vision represents an insightful perspective rather than a visual impairment.

A person possessing microscopic vision can easily discern intricate aspects of their field of expertise. However, this heightened perception comes at a cost. There is a trade-off with overlooking other important aspects of the development context. After mapping these professional biases (shown initially in Figure 9) developers questioned their "microscopic vision." Although not using that technical term, they comment on how a change of perspective can aid them in their initial goal of contributing to healthcare services:

> Developer B: “So maybe… we should also look at the structure of public health and how the model influences what is done there. If we try to get out of our comfort zone to foresee these possibilities, maybe we can truly help people, and create models that can be reproducible… I question what we did in the paper, which literally was improving the performance of that 1% we needed…”
>
> Developer A: “I even question the fact that if we are all primarily engineers in a team, but we are working on medical stuff, then you have to make sure you assess the representativeness of that domain, to identify deficiencies. It is the talking, more than just a list or writing a protocol, it is this discussion that brings up new things.”

By metaphorically looking into the microscope, individuals focusing intensely on specific details can potentially miss aspects that are directly influencing their decision-making, as seen in the case study. Hence, achieving a balanced perspective requires individuals to momentarily set aside their microscopic vision and consider the broader context to gain a comprehensive understanding of the overall situation.

But what can be done to counteract the influence of this microscopic vision? According to Davis (1998), a professional with microscopic vision “need only cease using his special powers to see what others see. He need only look up from the microscope.” (p.122-23) In this sense, Davis points out that this is not an unavoidable condition, but a burden that is acquired through specialisation. Thus, for Davis, although microscopic vision is a sort of power, it has its price: “You cannot both look into the microscope and see what you would see if you did not.” (Davis, 1998, p.123)

However, as shown by the developer team in the case study, they were able to have a “non-microscopic discussion” by following the BNA. This leads me to think that the Bias Network Approach could aid developers to “look up from the microscope,” without disregarding their necessary and valuable technical expertise. Thus, following Haraway (1988) and Draude et al. (2019), helping them recognise and understand their situated knowledges and adopting a perspective of strong objectivity (Harding, 1992, 2015).

Although I consider that Davis’ metaphor does highlight an issue that affects engineers, it is worth noticing that claiming that all engineers suffer systematically from this microscopic vision might be reductionist. For example, professional bodies for engineers trained in environmental or civil engineering, do incorporate certain considerations for context and the importance of risk assessments and broader impact, as this is part of their training.

Hence, I do not wish to generalise this claim or stigmatise a certain group of professionals. Still, the developer team's comments about their professional biases in this case study are similar to concerns raised in the AI ethics literature about technocentric views for AI solutions, which means focusing on performance, metrics, and outcomes without always taking into account the factors that affect how they are developed.

Likewise, the gynaecological field —as shown in the discussion above, is particularly problematic when it comes to having consistent diagnosis training data, given the biases of medical professionals towards a certain type of language or the avoidance of clinical terms by midwives. Therefore, I will broaden the considerations of this microscopic vision metaphor to professional biases in general (not just engineers), considering professional traits, tendencies, or acquired practices influencing decision-making and consequent behaviour, a type of "professional deformation."

A complementary concept to microscopic vision is professional deformation. Polyakova (2014) uses this term to talk about a phenomenon in professional ethics, where professionals of a particular discipline undergo a cognitive transformation that makes them develop thinking and behavioural patterns strongly influenced by their professional training and their daily practice standards:

> "Professional deformation of the personality is a sum of changes in cognitive structures, personality traits, behavior, methods of communication, stereotypes of perception, one's nature, values, and such caused by the execution of one's professional duties." (Polyakova, 2014, p.280)

Although different aspects can characterise professional deformation, there are a few I will highlight here, considering the context of AI. First, professional training involves shaping perspectives to fit the technical skills and problem-solving approaches specific to the discipline. This can, of course, translate into the technocentric practices highlighted by the developer team in the case study and discussed in previous chapters. Second, it can instil a narrow perspective on how to approach interdisciplinary problems. In the case study, this was reflected in how the developer team discussed fairness, mainly through metrics and without much consideration for broader context:

> Developer B: “Beyond performance metrics or accuracy adjustments we did not think about “fairness” in general or in a broader sense.”

The developer team characterised their professional biases as one of the main sources of impact for their decision-making and problem-solving processes when developing their model. They prioritised technical accuracy and optimisation of the model’s performance metrics, through decisions about sampling and technical fairness definitions. Focused on achieving optimal performance of their model, they prioritised achieving high accuracy rates, disregarding issues about representativeness —or at least not being aware of the issues to report and reflect on them as part of their developing process —such as the discussion about representativeness related to medical data or the societal biases influencing the gynaecological field.

Accordingly, I believe there is room to argue that professional bias should be more openly discussed in the AI bias literature, as a type of bias that transcends the typical category of societal, cognitive, and technical biases, and instead puts light into the practices and methodological limitations AI developers acquire through their professional exercise. This becomes particularly relevant when discussing sociotechnical views or solutions, which are based on interdisciplinarity and contextual requirements.

4.3.3 Conclusions from the case study: assessing the Bias Network Approach.

Amongst the main findings and the insights provided by the developer team in applying the Bias Network Approach, some key elements show the potential its potential to offer a sociotechnical complement to existing ethical guidance and frameworks to deal with bias in AI.

1. **It is a people's-based approach:** A sociotechnical approach, in its essence, implies fostering dialogue and collaboration to achieve the goal of contextualising AI. Hence, the BNA can function as a translational bridge that brings technical and societal considerations into the AI developers' domain of practice. By prioritising the dialogue amongst developers, offering visualisation tools, and fostering the consideration of external and contextual constraints affecting the developers' decision-making process, the approach offers an accessible and collaborative sociotechnical alternative.

Furthermore, in this pilot, the approach was implemented with the active involvement of an external expert group, stimulating, and facilitating the process. The developer team highlighted this dimension as a significant one. As mentioned by developer A: "[…] the thing that really helps and nourishes the process is that talk with the team members is the guided conversation […]". Overall, the approach encompasses diverse inputs that are created and implemented organically through dialogue, including a visual mapping, all fostering ethical awareness.

2. **It can facilitate transparency and explication:** The developers observed that the visual representation of interconnectedness elements within the BNA could serve as a valuable tool to explain their decision-making processes and the intricate interplay of various influencing factors. They recognised the potential of this approach to offer a comprehensible rationale for their choices, thereby enhancing their ability to communicate these analyses effectively and, more importantly, reflect on them. This communication could extend to scholarly publications, as mentioned above, but also promote further engagement with other stakeholders, as noticed by one of the developers:

> Developer B: "Like, I know our solution is not perfect but is magnitudes better than what the Ministry [of Health] has. So maybe having more consideration for their context can help us show why they need this and how this can help. Developer C gave some training to a few experts at the statistics department in the Ministry of Health, mainly Python. But the rest of the people involved in implementing the model had no idea about technical or other social benefits. So, the interpretation of any

> results or benefits is super unbalanced. We should have some sort of online feedback to report errors or worries. We did not consider that.”

Hence, the process of making this network that maps influential factors and biases in the development process helps both the developer team to gain insights about the ethical dimensions of their project, and the community, benefitting from having further information about the decisions made by the developer team, that can be communicated easily when contextual factors can explain pitfalls and limitations influencing the work of the developer team.

Thus, I identify two levels in which the Bias Network Approach can facilitate the ethical assessment of AI bias: (i) by offering parallel modes of assessment: visualisation, tracking and mapping, discussion, and articulating these with guidelines and frameworks; and (ii) by being practical, meaning that it does not only consider context or external elements conceptually, but it also connects this identification to the professional practice of the AI developers.

Aside from these foreseen benefits from the pilot, it is crucial to stress an important pitfall of the approach: its dependence on external experts for the initiation of prompting processes. For now, the more tangible benefits are rooted in the network aspect of the approach, which promotes contextualisation for a sociotechnical view of AI development.

There are also other limitations worth considering. One limitation is the generalisability of the case study. The specific results of the case study are not generalisable. However, there are enough distinctive patterns in these findings that allow me to see universal benefits to the approach, particularly regarding the benefits for AI developers to contextualise and improve their ethical decision- making.

It is also worth noticing that, while the BNA promotes interdisciplinary collaboration as a strength, operationalising this in practice could pose significant challenges. Differences in terminology, methodology, and professional cultures between disciplines (e.g., engineering, ethics, social sciences, medicine, and others) can hinder effective collaboration and integration of diverse perspectives. It is worth considering that the prompter team should at least have the interdisciplinary experience to overcome this.

Furthermore, regarding the pilot case study discussed in this thesis, the developer team was in a relatively advantageous position concerning documentation. However, this could readily become a significant challenge for other retrospective interventions or cases where documentation is insufficient or unavailable. The developer team maintained a well-documented record of certain decision-making processes (predominantly informed by technical criteria), alongside datasets, variables, model testing procedures, outcomes, and feedback specifically for research purposes.

Nevertheless, if the developer team collaborates with third parties to acquire data and encounters limitations in their capacity to trace its origins, or if they fail to comprehensively document their methodological decisions, this could substantially hinder the efficacy and/or depth of the insights provided by the BNA intervention.

Additional limitations could include discrepancies in data provenance, such as datasets being modified or merged without adequate records, or the use of proprietary datasets from external sources, where legal or contractual restrictions prevent detailed auditing. Another issue might arise from the lack of version control in model development, making it difficult to reconstruct the exact configuration of algorithms at different stages of development. Furthermore, if the data processing pipelines include steps that are poorly annotated or executed using opaque tools, it may obstruct the ability to identify bias, assess fairness, or trace decisions back to their technical or conceptual origins.

There could also be challenges associated with the absence of stakeholder involvement or insufficient documentation of stakeholder feedback during the development process, particularly in systems intended for high-stakes applications. This could result in gaps in understanding the social, cultural, or ethical considerations embedded in the decision-making process. Finally, resource constraints, such as limited personnel, time, or funding, could lead to incomplete documentation and tracking, thereby reducing the capacity for a BNA intervention to effectively uncover systemic issues or propose actionable improvements.

To overcome these limitations, as well as ponder other potential hurdles discussed in previous chapters, such as implementing the BNA at different stages of development or evaluating how demanding it can result in terms of human resources, more case studies and implementations will need to be examined.

Future research should also investigate the scalability of the BNA when applied to larger, more complex AI systems, particularly those involving extensive collaborations across multiple organisations or jurisdictions. This might include examining how to streamline data collection and annotation practices, ensuring traceability in cases where resources are constrained. Moreover, evaluating the feasibility of integrating BNA principles into existing workflows and standards for AI development could help identify best practices for embedding ethical oversight without overwhelming project timelines or budgets.

Finally, the examination of real-world applications should consider the broader organisational and cultural factors that influence the adoption of frameworks like the BNA. For instance, resistance to transparency or accountability due to competitive or proprietary concerns might necessitate specific policy recommendations or incentives to encourage more open and systematic documentation practices. By addressing these challenges through iterative implementation and refinement across diverse contexts, the BNA can be strengthened as a reflective ethical intervention for promoting responsible, and effective AI development.

Chapter 5: Guidance for Developers and Prompters.

In Chapter 4, I talked about my idea for a Bias Network Approach, which is a sociotechnical way to help AI developers think critically about biases. This would help them see things from a broader context and become more aware of the different factors, choices, and limitations that affect the role of biases in AI development.

In this Chapter, I will first provide some preliminary guidance for developers and prompters to apply the BNA in section 5.1 and then conclude by discussing the advantages of applying this approach in section 5.2.

5.1 Applying the Bias Network Approach (BNA)

As described in Chapter 4, this approach requires a prompter team, which helps developer teams map the connections and influences guiding their decision-making, bridging technical and societal considerations.

Here, I will offer concrete guidance on how one can apply this approach, considering that many of these requirements will be contextual, so they can be adapted to the needs of each team and project. In the future, this guidance will also be informed by further case studies and more specific distinctions could be made for different types of projects, stages of development, etc.

To facilitate this explanation, I will divide the guidelines into the three stages used to implement the BNA: the preliminary stage (preparing for the BNA intervention), the intervention stage (interaction between the prompter and developer teams), and the follow-up stage (post-intervention queries).

5.1.1 Preliminary stage.

The preliminary stage consists of requirements for both the developer and the prompter team, that are necessary for both to acquire valuable information about the AI project to prepare for the intervention stage.

- For the developer team.

1. Identifying the profile of the developer team.

The developer team should characterise the background of the developers who are working on the project. They should know if they have worked on interdisciplinary teams before or ask themselves questions about their "sociotechnical experience". For example, "Do any of us have experience with projects that have a high social impact?" "What are our main disciplinary specialisations?" "Have we applied ethical evaluations or assessments to our work before?" —This profiling should be performed by the developer team before the interview with the prompters (ideally) or it can be arranged to be part of the intervention as well. Overall, the purpose is to discuss and for both the developer and prompter teams to be aware of weaknesses and strengths the developers might have both individually and collectively.

2. Define the scope and principles guiding the project.

Clearly outline the goals and objectives of the AI project. This involves specifying the problem that will be solved (or tackled), the data that will be used (or was used if it is a retrospective intervention), and the intended application of the AI system (acknowledging the context and relevant stakeholders). Aside from defining the scope, it is also encouraged that there is an identification of the guiding ethical principles for the project. The developer team should be able to point out which principles they have considered (e.g., in the case study, the developers only talked about fairness, and focused on metrics and accuracy trade-offs), and if no principles have been considered, also bring this forward at the intervention interview.

3. Provide a summary of the project and technical information.

The developer team will have to create a document or a set of documents with relevant information about the project for the prompter team to peruse. If it is a retrospective intervention like in the case study, any publications, tests, or datasets can be shared to complement a summary description of the project. This summary can also involve information about specific decisions they have made or have considered, such as using a particular tool for bias mitigation, fairness definition, or any steps for data selection.

- For the prompter team.

1. Get to know the project

This step involves gaining as deep of an understanding of the AI project as possible. For this, the prompters should familiarise themselves with the project's goals and objectives for each development stage. Special attention should be given to identifying the sources of data that will be used in the project. For example, questioning if there is any reported information about where the data comes from, its quality, and any potential biases that may exist in the data, as well as thinking about possible biases the prompters consider could appear in the development process (and that the developer team could miss).

The prompter team should get information from the developers about the project's scope. If something needs clarifying, prompters could decide to either ask for further information before the intervention or ask during the interview. For example, are the specific tasks or decisions that the AI system will be involved in clear? Understanding the boundaries of the project is essential. The prompters should also consider the broader context in which the AI system will be deployed. How will it be used in the real world? Understanding the application of the AI system is critical for assessing its potential societal impact and, therefore, visualise biases that can interfere with the goals of the project and identify the principles that should guide the development process.

2. Prepare the prompts.

To prompt the developer team, a list of common biases for each developing stage can be used (it could be prepared by the prompters, or they could use the one provided in the literature review here or other reviews). Prompters could also mention some relevant biases concerning the project's goal and context of application, as well as other specific biases that could apply to the configuration of the developer team and their disciplinary background, e.g., highly technical hence the potential to disregard societal aspects, or very interdisciplinary, which could contribute to a lack of consensus in respect to certain delimitations for key concepts like fairness or even bias.

The prompts should have the following characteristics:

• Relevance: each prompt should be relevant to the specific AI project at hand. Consider how the biases or sociotechnical factors mentioned in the prompts relate to the project's goals, data, and intended applications. Examples: "How could biases in the data sources used to impact the accuracy of the AI model in the context of healthcare diagnostics?" or "Considering the project's goal of automated loan approval, what potential biases could arise in the machine's decision-making process?" —the important aspect is to emphasise the context of the application of the AI system and the factors influencing its development, to integrate any insights to the development context.

• Open-Ended Nature: the prompts should be designed to encourage open-ended responses. Avoid yes/no questions and instead ask questions that prompt deep reflection and discussion, and that stimulate the developers to think about their own decisions and perspectives. "What ethical considerations arise when deciding which features to include or exclude from the algorithm?" "Which potential biases could affect the fairness metric you intend to use?" (Perhaps offer some options if they feel stuck).

• Multidimensional approach: the prompts should cover a range of dimensions, including technical, ethical, and societal aspects, ensuring a comprehensive exploration of potential biases and issues. For example, they could ask: "Apart from technical accuracy, what ethical and societal factors should be considered when evaluating the success of your AI system?" "How do technical challenges (give a concrete example, ideally brought by them) intersect with broader societal issues (provide an example to connect them), and how can we address these intersections?" "What ethical principles should guide the decision-making process when selecting data sources for the AI system (considering other elements discussed)?". The prompts should connect the different dimensions of inquiry and guide the developers to make this connection.

• Sensitivity to disciplinary backgrounds: the prompts should be sensitive to the disciplinary backgrounds of the developer team. If the team is highly technical, include prompts that encourage consideration of societal aspects. Conversely, if the team is more interdisciplinary, address prompts that help build consensus on key concepts like fairness and bias. This will be a consideration that is harder to universalise. Overall, the important aspect is to consider the potential conceptual gaps that come from interdisciplinary talk. Make sure you define your terms and feel free to also ask them to define theirs.

- Flexibility: there must be flexibility in the prompts, contextual improvisation is encouraged. Unlike studies where you would require the interview to follow a specific script to ensure consistency, in this case, the goal is to encourage developers to engage in critical ethical thinking. Depending on the discussions and emerging issues, prompters should adapt and expand on certain questions. For example, in the case study, as prompters, we encountered the developer team's worries about their professional biases, so we explored this by asking how they identified this as a limitation or how this affected their decision- making in relation to fairness criteria. Although they did not use the concept of "professional bias" we characterised it based on the descriptions provided by the developers.

- Avoid leading questions: it is important that prompters do not suggest specific answers or biases without being brought up by the developer team. This is different from presenting a set of possible biases for them to consider and ponder. The prompter interventions should aim to facilitate the discussion amongst the team.

Hence, the role of the prompters is to stimulate collaboration, dialogue, and a reflective stance amongst the developers. Ideally, the mapping during the intervention should only include the findings of the developer team. However, if the prompters see that a particular bias or principle was not brought up in their conversation and could benefit the development process, they could add it as a suggestion in an alternative version of the network map.

The same could apply if there is a pressing issue that was not brought up by the developers, but the prompters consider critical. Notwithstanding, the intervention experience should guide the developer and not evaluate their answers.

This is crucial, as the BNA is not to be perceived as an evaluative instance, but rather as a supportive and reflexive stage of their project development. By preparing prompts with these characteristics, prompters can guide meaningful discussions and facilitate a thorough exploration of sociotechnical aspects and map elements, actors, factors, decisions, principles, biases, and limitations encountered by the developer team.

5.1.2 Intervention stage.

The intervention stage is the main core interaction between the developer and the prompter team. Both teams should discuss and reflect on biases from a sociotechnical perspective. The prompters will help map the network of influences identified by the developer team.

- For the prompters.

1. Conduct semi-structured interviews.

The prompters will conduct semi-structured interviews with the developer team to discuss and identify relevant elements and factors related to the project. These interviews will serve as the main way to engage with the developer team. This instance should be based on dialogue and collaboration amongst the members of the developer team, making them the protagonists. The structure of the interview should be based on the prompts and project details prepared by the prompter team. But, as mentioned above, it should allow for flexibility and be open-ended.

2. Illustrate the network map.

Prompters will use the network map to help developers visualise potential connections between biases and other relevant factors. The map should be built around the core AI pipeline structure (Figure 1), making it accessible for AI developers. It should include colour-coded categories for visualising biases and their interactions considering decision-making criteria, material limitations, and potential sources of bias. The map should be explicit, showing the specific connections and relevance amongst the nodes of the colour-coded categories used. The illustration can be replicated using different design tools or even drawn manually to make the session more interactive.[15]

3. Encourage the developers to document their developing process (after the BNA intervention).

Part of the intention of developing the BNA is not only to make sociotechnical integrations into AI development more accessible, but also to support documentation of the discussions, decisions, and changes made throughout the application of the BNA, thus serving as a valuable accountability resource. Proper documentation facilitates transparency about the AI development process by, for example, showing a historical account of the decisions made, the actions taken, and the rationale behind them. This transparency is vital for holding developers, accountable for their choices—when relevant. The BNA not only helps developers keep track of this information but also offers further insight into the connection of different relevant factors that justify the ethical decisions behind their sociotechnical choices. Developers are encouraged to keep mapping decisions, biases, and other elements they encounter after an initial intervention.

5.1.3 Follow-up stage.

The follow-up stage refers to any instance following an intervention stage intended to update or check if the findings of the first intervention have been adequately incorporated into the design and development choices of the developer team.

- For the developer team.

Contact the prompters: if needed, the developer team can contact the prompter team to clarify doubts or even request a second intervention. If, as the project advances, further aspects, different from those identified in the intervention, keep arising, it is recommended that a second intervention is conducted.

[15] Should you wish to use the same visualisation I used in this work, a template for the bias network is available here: BNA mapping template. A GitHub repository with and interactive platform to create the network is currently being developed.

The developer team is advised to keep updated documentation during development about factors, elements, and decisions based on the mapping they did in the intervention. If there is a second intervention, this documentation should be shared with the prompter team to prepare for the second interview or any other consultation.

5.1.4 Timelines for applying the Bias Network Approach.

Considering the first test was the pilot case study, I am only able to provide estimations for the timing requirements for both the prompter and developer teams. These estimates, however, can vary depending on the contextual requirements of each project.

The pilot case study presented in Chapter 4, was a retrospective intervention. For this intervention, the total time invested by the prompter team was around 8 hours for each member, including both the preliminary and the intervention stages. This was possible because each member of the prompter team worked on different tasks simultaneously. If the prompter team decided to divide times differently, this average of invested hours could change.

In the case of the developer team, the initial time investment was related to preparing the case description, goals, objectives, and ethical concerns. Then, for the intervention, the developer team invested around 2 hours, and 1 more hour to do some follow-ups, mainly to get further insights about the experience (for research purposes to report about the case study). For the interventions and preparing the information as prompters, we spent around 8 hours total each (24 hours total).

After the first intervention, it is expected for the developer team to spend time applying the network approach as part of their documentation process, checking on decisions and factors influencing their development.

This, however, should be considered as part of their developing tasks, because if they wish to follow a sociotechnical view, then this is integrated into their continuous practice. In this sense, continuing to apply the BNA, would not be a “side task” from their main developing practices.

5.2 Why is the Bias Network Approach an advantage?

In the last two chapters, I have presented an approach to help AI developers integrate a sociotechnical view to address biases in AI development. Offering a structured, collaborative, and interdisciplinary example of how to apply the BNA through an intervention like in the case study, where developers were encouraged to uncover, monitor, and visualise the key bias-related factors that impact their ethical decisions made during development.

When I introduced the BNA at various conferences, I encountered a common concern, especially during seminar presentations with AI researchers and representatives from the private sector: the human resources required to implement the approach. This potential concern, however, I do not consider it discouraging, mainly because the long-term advantages of adopting the BNA could far outweigh the initial investment.

Imagine an institution whose main production is focused on AI services; therefore, it has a variety of simultaneous projects in different sectors, e.g., education, healthcare, and climate. This institution wants to align its development process with a sociotechnical perspective like the BNA. If this institution decides to implement the BNA, they will need to assemble a prompter team to support their developer teams. While this may involve seeking external expertise (or gathering an internal expert group), these interventions do not demand an excessive allocation of resources when compared to the substantial technical and ethical benefits they can deliver. Adopting a BNA can improve ethical engagement and promote self-responsibility (this will be discussed in the next chapter), produce visualisation and documentation that justifies decision-making, as well as adopt a preventive approach to biases (foreseeing their influence within the network).

In essence, the initial commitment to building a prompter team and implementing the BNA is a small step when viewed against the significant gains. The long-term payoffs, in terms of more responsible AI development, can make the investment well worth it. Accordingly, I will show three potential benefits of applying the BNA to an AI project, which could make the BNA look like a necessity.

5.2.1 Avoiding superficial technical fixes.

When using the network approach to spot potential biases, identifying key factors at play and the decisions that revolve around them, there is a multiplicity of issues being addressed. There is a cautionary measure being placed because this can aid in preventing the reoccurrence of problems, that otherwise could require constant technical patch-ups.

Imagine you, as a developer, find a bias during the data collection phase and take the necessary steps to rectify it. What the network approach does is help developers see the bigger picture. Hence, they can grasp how this bias might affect not only the initial data but also ripple through subsequent stages of model development, validation, and even how the AI system is used after deployment.

This comprehensive view prompts developers to make systematic changes rather than relying solely on isolated technical fixes. This can be translated into you, the developer, still making the same decision to fix that data collection bias.

However, the difference, is that before technically fixing it, you have mapped the bias. You question its origin; you see if it has any connection with the sampling made by your team or an external one. You explore if there is any connection between that data and other data you will be using, making sure the same bias is not replicated, etc. In simple words, you think beyond a specific instance of bias.

With this newfound awareness, developers adopt a more comprehensive approach. They are prompted to rethink their (or others) data collection methods, now informed by the ethical insights offered by the BNA. As a response, you might want to revamp the validation and testing process, for example.

The important change is in how you understand bias. Now your actions are not defaulting into technocentric fixes and, instead, they instigate systemic changes within the AI development process.

Addressing biases is no longer an isolated rectification of an immediate bias; it is about crafting a robust and situated standpoint, changing the way the developer team understands bias issues from a sociotechnical view. Accordingly, the BNA promotes developers to become proactive actors that conceptualise biases as integrated within a network of influences. This approach ensures that future projects from the same team, institution, or even the wider community can benefit from the wisdom gained by mapping these interactions, fostering a culture of transparency and constant improvement.

In summary, the BNA sheds light on the interconnected elements and decisions in AI development, as well as recognising how biases can influence and interact with each other, discouraging the adoption of the views criticised as the "problems of bias" Chapter 1. It promotes a forward-thinking, systemic approach to address biases and other challenges comprehensively.

5.2.2 Benefits of documenting decision-making with the BNA.

Documenting the development team's ethical decision-making offers a transparent roadmap of how and why specific considerations were put into practice. This documentation is critical because it offers a dual benefit.

First, ensure rigorous ethical standards, providing a record of the decisions and limitations considered to deal with bias. Second, this documentation lays a foundation for future ethical considerations, as previous decisions offer valuable insights from lessons learned in the past.

Moreover, around the world, different legal and regulatory requirements are governing aspects such as data privacy, bias mitigation, and considerations for ethical risks and negative impact of AI. Thorough documentation is crucial to demonstrating compliance with these regulations. Thus, not only minimises the legal risks faced by organisations but also equips developer teams to adapt to new ethical and regulatory requirements as they emerge. For instance, consider the example of the Chilean government (given that the case study was conducted in this context).

Chile is in the process of developing its legislation concerning AI and the use of sensitive personal data. These laws are centred around the assessment of risk. If the use of sensitive personal data or an AI system presents a high level of risk (think credit applications, healthcare, or education), those responsible for the project are mandated to provide a systematic evaluation of their development process.

This evaluation must meet specific requirements. Documenting the treatment of data is like painting a clear picture of how information is handled, including a statement of the aim and purpose behind this data treatment. This, in turn, requires an input data management plan, considering the evaluation of potential biases and methods to spot gaps or data deficiencies, along with strategies to address them.

Thus, documentation is not just an administrative task; it is also a powerful means of communication. Documentation of development processes, as the one suggested with the BNA, allows different stakeholders —like other developers, interdisciplinary experts, users, and regulators— to understand the ethical considerations and decision-making that drive an AI project. Open and effective communication is crucial to promote much-needed interdisciplinary collaborations sociotechnical views encourage and assist with important tasks such as post- deployment monitoring, knowledge transfer, and even building trust among the public.

In a nutshell, documenting the AI development process and ethical decision-making should not be a mere formality. It is a crucial practice that fuels ongoing ethical improvement. The documentation provided with the BNA, combined with other available documentation practices, makes the process of AI development more robust because developers do not capture isolated steps; the network mapping is a tool for unveiling the intricate web of sociotechnical choices and factors affecting the AI development journey. Hence, the focus of the BNA is not only to look into what was done but also on why and how certain decisions were reached.

This brings a richer context to the discussion of the ethical and sociotechnical aspects of AI development. By visually connecting the dots between various elements

—like data sources, modelling techniques, and decision-making instances— it becomes easier to follow the evolution of the project's ethical considerations and foresee the potential consequences of those decisions.

Moreover, the BNA encourages developers to think about the bigger picture at every stage of development, making the BNA an active resource for ongoing ethical engagement.

Hence, when combining the documentation provided by the BNA with other existing means of documentation, you get a comprehensive record of AI development: the "what" plus the "why" and the "how" of ethical and sociotechnical decision- making in bias management. This synergy in the BNA strengthens transparency, accountability, and continuous ethical improvement in AI development.

5.3 Conclusions for this Chapter.

The BNA can be a solid tool to integrate sociotechnical perspectives into ethical evaluations in AI development. Despite biases being recently recognised in the literature as related to one another (Schwartz et al., 2022), there is a lack of systematic efforts to establish an explicit relation between specific biases and contextualise them within the context of AI projects.

In the discussed pilot case study in Chapter 4, developers manifested how the visualisation map allowed them to reflect on their decision-making process, but also how this could help them communicate this process. For this, I believe it is a reasonable expectation that in further case studies the visualisation aspect of the BNA will continue to facilitate transparent and accessible explanations of the ethical considerations made by developer teams.

As AI advances and becomes integrated into society, the relevance of comprehensive ethical evaluation and responsible AI development cannot be overstated. Adopting an approach like the BNA has the potential to play a crucial role in supporting developers, and even organisations and other stakeholders. Overall, it can help us shift into perceiving AI ethics as an active and robust way to engage with intersectional understandings of AI and development practices.

In the next chapter, I will examine the responsibilities we can attribute to the different actors involved when the BNA is adopted, including developers, prompters, and institutions.

Chapter 6: Responsibilities

In my thesis, I have developed a detailed sociotechnical approach that aims to enhance ethical decision-making and the evaluation of biases in AI development. This approach is designed to integrate ethical considerations seamlessly into the AI development process, and it has been designed to support developers willing to integrate ethics into their practices. Hence, this approach is not intended to deter companies like "Evil Corp." Such entities are unlikely to voluntarily adopt this approach without legal incentives or coercions. My proposal is better suited for organisations, professionals, research teams, and developers who are ethically inclined but lack the know-how to apply a sociotechnical approach effectively, facilitating a more profound integration of ethical considerations in their work, and enabling them to make better methodological choices about bias.

However, there may be compelling reasons for even the likes of Evil Corp to consider adopting an approach like mine, which I will elaborate on in the final conclusions of this chapter.

To introduce the chapter, in section 6.1, I will give a brief description of responsibility as a moral obligation in AI based on Tollon's (2022) proposal, as I will then relate this to responsibilities that come from adopting the BNA. Then, in section 6.2, I will first distinguish between forward-looking and backward-looking responsibility and then, I will connect these with the BNA as a promoter of active responsibility and notions of the ethical agency of AI developers, based on a study by Griffins et al. (2023). Then in 6.3, I will briefly comment on some of the responsibilities of different actors involved in adopting the BNA, on an individual level (developers and prompters) and an institutional level (companies, educational institutions, and professional bodies).

6.1 Responsibility as a moral obligation in AI.

In the general discussion about responsibility, there is one concept worth emphasising here: responsibility as a moral obligation.

Fabio Tollon (2022) defines responsibility as a moral obligation as:

> "a responsibility for future states of affairs and is concerned with the active promotion of certain societal goals, and the responsibility of agents to align what they do with these goals (Santoni de Sio & Mecacci, 2021) We must take seriously our obligation to ensure that the decisions we make today help in the pursuit of a better tomorrow." (Tollon, 2022, p.308)

Tollon argues that, to ensure that "better tomorrow," it is insufficient to simply consider the future impact of our decisions, it is necessary to have active obligations that can push society in that direction.

More specifically, for agents to be responsible for these future states, according to Tollon, they "ought to be in some sense under their control" (Ibid), for us to claim they are responsible for them. Therefore, there is a moral obligation that can be predicated on the control agents should have over the creation and deployment of AI systems.

A problem with this, however, is that this can present an important challenge considering the capabilities of AI's emergent behaviour, i.e., its capacity to learn post-deployment, and the experimental nature of AI systems, which can sometimes mean that the goals that were programmed into these systems might be achieved in other ways outside of human control, thus creating a responsibility gap.[16]

The important aspect I wish to highlight here, without entering the responsibility gap debate, is what Tollon says about a possible way to overcome this gap.

[16] The problem of different responsibility gaps in AI is quite extensive. I do not wish to enter this debate here. I will, however, highlight some important things to take into consideration. The concept was originally introduced in 2004 by Andreas Matthias in the specific context of advanced learning machines, where he argues that intelligent systems capable of learning from their interactions and environment become so complex and unpredictable that human control over them diminishes significantly which creates the challenge of assigning responsibility. More recently, the issue of responsibility gaps has been more broadly discussed. Risks associated with AI and the extent to which individuals should be responsible for AI's actions have become a central and more nuanced question in the field of AI ethics (Braun et al., 2021; Coeckelbergh, 2020). Discussions are now involving complex socio-technical systems, which include less autonomous AI and intricate networks of human agents and technical systems, which can also lead to responsibility gaps. Thus, as argued by Sio and Mecacci (2021) a more comprehensive understanding of responsibility gaps, considers factors beyond the AI system's autonomy, a broader consideration that is crucial for devising more effective solutions to these ethical challenges. And, as mentioned here, some of these factors have to do with the AI developers' decisions.

In short, he argues that such a gap is not unsurmountable and that there are ways to overcome this issue:

> "[…] designers and developers ought to regularly check that the AI in question is performing its task in a way that is aligned with various socially desirable values (respect for human rights, equality, sustainability, etc.). This would involve understanding the specific context in which the AI is embedded, as well as how the agents interacting with it understand it, and how it affects the communities and groups within its range of influence." (Tollon, 2022, p.316)

To achieve this, Tollon highlights the importance of developing a hermeneutical approach, which implies not just looking at the possible consequences of technology, but also paying attention to how the technology is being understood. This process of understanding AI development is what Tollon highlights as an iterative process i.e., the hermeneutic circle:

> "Once we take the time to understand the social meaning of a technology we do not come back to our original starting position. Rather, the process of uncovering meaning itself creates a kind of spiral, whereby new inputs are interpreted by society in a number of ways and come to influence our understanding of the technology in question." (Tollon, 2022, p.315)

Following this hermeneutic approach, Tollon points out the importance of focusing on the process of decision-making. Some risks can be assessed and prevented during AI development, i.e. adopting a responsible research culture. However, adopting this stance does not come without challenges. Tollon, emphasises that the adoption of moral responsibility linked to decision-making is not straightforward, because engineers in isolation might not be able to fulfil these moral obligations "without education and input from researchers in the social sciences." (Tollon, 2022, p.316)

What Tollon proposes is to develop inter- and trans-disciplinary work:

> "[…] so that the given societal meaning of the system can be uncovered. Such a process demands a diverse and pluralistic approach to technological assessment. Additionally, it might seem excessively onerous that programmers or engineers have to undertake such a hermeneutic analysis. This is especially concerning if we reflect on the gap between theory and practice that is operative in the AI ethics debate at present (Morley et al., 2021)." (Tollon, 2022, p.316)

Overall, there are at least three important aspects presented by Tollon that are relevant to the responsibility discussion regarding the BNA.

First, an important way to attribute responsibility is based on the sociotechnical understanding of AI systems. Second, AI developers should understand the societal significance and the interactions of the technology, focusing on the decision-making process that justifies how they create the system. Third, forward-looking moral obligations as argued by Tollon can be challenging but, if a sociotechnical standpoint is adopted, this can help overcome complications in responsibility ascriptions.

The responsibilities that come with adopting the BNA align with the arguments proposed by Tollon. I believe the BNA as a methodology is complementary to the iterative aspect of the reflective hermeneutic spiral, which combined with a sociotechnical standpoint, actively helps developers incorporate the social meaning into AI development and, therefore, can contribute to clarifying responsibilities.

In what follows I will show how forward- and backward-looking responsibilities relate specifically to the BNA.

6.2 Responsibility in AI and its relation to the BNA.

The BNA provides novel insight to address biases, which I argue can enhance the sense of responsibility among AI developers. How? It involves mapping the interconnections among key elements such as data sources, influential factors, and decision-making. This mapping creates a comprehensive picture of potential bias origins and pathways, equipping developers with the knowledge to identify and mitigate biases proactively including their own and external limitations, to deal with them. It also facilitates transparency with the developer team justifying their decisions with a clear understanding of the biases and influences being addressed.

Thus, the bias network approach can serve as an enabler for responsibility attribution concerning the design and development of AI systems. By mapping out the biases present in an AI system's decision-making process, this approach can help to visualise, document, and intervene in ways that clarify how decisions are made and how influences are considered.

By representing the flow of information and decision- making graphically, it is easier to highlight areas where biases could potentially influence outcomes and the developers' or users' decision-making. This makes it easier for all stakeholders to understand the system's operations and the potential areas where biases could affect decision-making.

Furthermore, the process of building a bias network requires thorough documentation of the AI system's design, including data sources, algorithms, and decision-making pathways. This documentation can serve as a record that explains how the AI was intended to function (expected use and deployment), which also involves the considerations taken to minimise or identify bias, and the responsibilities of different actors in the development and maintenance of the system (overall reducing risks).

Therefore, with a clearer understanding of where biases may occur, stakeholders can implement targeted interventions to mitigate these biases or their effects. This could involve retraining AI models with more balanced data, adjusting algorithms, or changing how data is collected. It also allows for the establishment of oversight mechanisms that monitor for biased outcomes and adjust the system accordingly.

Thus, the Bias Network Approach can help create a more transparent and accountable development process by providing a structured way to understand, document, and address biases. It can also facilitate a clearer attribution of responsibility by showing how different components and actors within the AI system contribute to its outputs.

But before mentioning these specific responsibilities in relation to the adoption of the BNA, I will examine backward- and forward-looking responsibility more broadly as part of the moral responsibility literature in applied philosophy, drawing from Ibo van de Poel's analysis (2011).

6.2.1 Distinguishing backward- and forward-looking responsibility.

To start this section, I will first make a primary distinction between causal responsibility and moral responsibility.

On the one hand, causal responsibility refers to the relationship between a cause and an event, hence an agent is causally responsible if an action they perform is the cause of said event. When concerned with causal attributions of responsibility, there is no moral judgment involved.

On the other hand, moral responsibility, which is the type of responsibility I will be concerned with here, evaluates if an agent is morally responsible for an outcome, hence judging if we can hold them accountable.

One determining factor, particularly based on the analysis in section 6.1 is the agent's control over their actions and state of affairs, knowing they acted with knowledge of the consequences. Thus "the powers and capacities that are required for moral responsibility are not identical with an agent's causal powers, so we cannot infer moral responsibility from an assignment of causal responsibility." (Talbert, 2023, para. 6)

Nevertheless, even though causal and moral responsibility are concerned with different aspects of responsibility, the former can inform the latter by establishing causal connections that can often be a requirement for moral evaluations, looking at who or what caused a particular state of affairs, and this is important for backward- looking responsibility. Talbert gives an example to illustrate this:

> "Suppose that S causes an explosion by flipping a switch: the fact that S had no reason to expect such a consequence from flipping the switch might call into question his moral responsibility (or at least his blameworthiness) for the explosion without altering his causal contribution to it." (Talbert, 2023, para. 6)

Now, when we talk about backward-looking responsibility, we refer to holding someone accountable for their actions. In simple words, backward-looking responsibility is concerned with an existing state of affairs, involving claims about who is responsible for the things that have already happened (van de Poel, 2011).

In this sense, establishing responsibility is a very direct relation between A (and agent) and X (a specific state of affairs), meaning that A is backwards-responsible for X. Responsibility can also be understood as a relational concept where "A is responsible for X to B" (B being a different agent).

Forward-looking responsibility, however, does not fit this basic relational form. In this case, an agent is responsible for something that could happen (or that she foresees as likely to happen). Van de Poel distinguishes this by explaining that forward-looking responsibility reflects:

> "the fact that we may have specific responsibilities to different people. Professionals like engineers, for example, have different responsibilities to their employer, to their colleagues, to their clients and to the public […] forward-looking responsibilities may arise from the specific relations we have with specific people (cf. Scheffler 1997). This is not to deny that we may also have responsibilities to ourselves or general responsibilities. In the case of forward-looking responsibility (2) [A is responsible for x to B] might then be understood as follows:
>
> A is forward-looking responsible for X to B means that A owes it to B to see to it that X." (van de Poel, 2011, p.41)

A simple example of this can be an AI developer (A) working for a client (B), to develop an AI system (X). In such a case, A (AI developer) is forward-looking responsible for X (developing the AI System) to B (client). This means that the AI developer owes it to the client to see that the AI system is developed according to the specifications agreed upon.

Now, consider a more specific scenario, where (B) is a healthcare provider, and

(X) is an AI assistant for patients scheduling appointments and providing information about medications or treatments. We could consider that (A) can have a responsibility that originates given AI system will have a significant impact on the patients' well- being. In this case, the AI developer (A) is also forward-looking responsible to the patients (C) who will be interacting with the AI system.

Furthermore, van de Poel's idea is that our forward-looking responsibilities can shape our backward-looking responsibilities. If you fail to develop (X), you could be held accountable for negative outcomes that result from that failure (depending on the circumstances). You might also find yourself accountable for a bad outcome, even if it was not your job to prevent it.

Van de Poel notices that one can have a backward- looking responsibility for a state of affairs "without having been forward-looking responsible for preventing that state of affairs" (p.50). Accordingly, he recognises that in the case of forward-looking responsibility, the emphasis is on being someone who acts responsibly (virtuous) and fulfils their duties (moral obligation). Whereas backward-looking responsibility focuses on explaining your actions (accountability) or if was your fault (blameworthiness).

6.2.2 Forward-looking and active responsibility, and the ethical agency of developers.

As discussed in 6.1 referencing Tollon, to fulfil certain responsibility requirements it is important to recognise that developers might not have the expertise to achieve the pluralistic approach required to have the necessary reflection to assess decision- making processes in hopes to avoid or prevent certain undesirable consequences or, at least, foresee potential issues and make responsible development decisions.

To support this lack of expertise, the BNA offers an interdisciplinary intervention that promotes responsible development. AI developers, possessing a deep understanding of the technology's inner workings, are in a privileged position to mitigate technical risks, but foreseeing societal risk is not necessarily part of their expertise. Therefore, an approach like the BNA can allow AI developers to fulfil that pivotal role in addressing ethical considerations.

Ultimately, the question of responsibility extends beyond the technical aspects of AI development encompassing the need for developers to engage with the wider context in which their technology will operate. I consider, therefore, that developers should be held accountable for the responsibility associated with this forward-looking moral obligation —when they have a certain control over the outcomes.

However, as stressed by Tallon (2022), this requires an inter- and trans-disciplinary approach that supports developers in adopting an active responsibility, to safeguard against negative impacts and promote positive ones (within their means). Hence, I suggest that if we want to ensure responsible AI development, adopting a sociotechnical approach like the BNA (or others) is necessary to fulfil this moral obligation.

With active responsibility, I refer to the proactive engagement of individuals and organisations in ensuring that the actions, decisions, and systems they create or manage are aligned with a sociotechnical understanding of AI development, in this case specifically related to assessing biases. It emphasises the duty to anticipate potential issues and to prevent harm, rather than simply reacting to adverse events after they have happened (mitigating damages).

These proactive measures, and understanding AI as complex sociotechnical systems, require a forward-looking approach that incorporates ethical considerations into the design and development of AI systems. In other words, by adopting the reflective and contextual standpoint required by the BNA, the requirements for an active and forward-looking responsibility are already being considered.

Connected to this notion of active responsibility is the ethical agency of AI developers. Griffin et al. (2023) explored this through semi-structured interviews with 40 developers, analysing the ethical aspects of their profession. The research highlighted three emergent themes: (i) ethics in the occupational ecosystem, (ii) ethical agency, and (iii) the characteristics of an ethical developer.

Regarding the first theme, ethics in the occupational ecosystem, there are three features discussed by the authors: personal aims, occupational morality, and technology's neutrality.

The first feature is personal aims. Developers declare themselves as well-intentioned, and all of them recognise the potential harms of AI, stressing that these are often caused by "a lack of knowledge or experience, [and] not intentional malfeasance." (Griffins et al., 2023, p.4) An interesting aspect of their responses highlighted by the authors is that "participants spoke in terms of what (or who) they were not rather than what (or who) they were.” For example, in one of the responses in the study, a developer said:

> “We don’t consider ourselves to be nefarious agents. It’s not that we have the pinnacle of high ethical standards, but we are also not evil geniuses sitting in our lair trying to figure out ways to hurt people.” (Griffins et al., 2023, p.4)

The interesting point here, I believe, is that the personal perception of the developer’s role might differ from the actual expectations or definitions of their responsibilities as AI developers. For example, consider an AI developer named Kai, who is working on a voice recognition system. Kai may personally believe he is being responsible because he consciously does not intend to create a biased system. He declares that: “He is not one of those developers who disregards ethical concerns.” The focus of such a statement is on distancing himself from negative traits rather than stating positive attributes such as, “I am a developer who prioritises creating unbiased AI systems by doing X.”

This self-perception could reflect a cautious approach to their identity as a developer, focusing on avoiding harm (by not intentionally provoking it) rather than actively promoting good, which might not fully align with the broader ethical expectations that developers should be proactively ensuring their development process is not negatively affected by biases. It suggests a potential gap between how AI developers see themselves and what is expected of them in terms of ethical standards in their profession.

The second feature is occupational morality and organising principle, where almost half of the participants declared that “development was neither ethical nor unethical” (ibid). In general, their perspectives responded to the field being new, so there is no real engagement with wondering if what they are doing is indeed ethical or unethical.

The third feature was brought up by the interviewees —as it was not part of the questions the authors prepared— and it highlighted the perceived neutrality of the technology. Interestingly, the authors highlight that “Practitioners’ belief in the neutrality of technology was strong conceptually but was shakier in practice.” (Griffins et al., 2023, p.5) This might reflect the lack of ethical training and conceptual tools to address the translational gap in understanding AI as a sociotechnical system, as discussed previously in this thesis.

The second emergent theme presented by Griffin et al. (2023) is ethical agency. The authors focused on two aspects of ethical agency, but a third category emerged in the interviews:

> "First, whether developers believed they could technically achieve what is being ethically asked of them. Second, whether they could intervene in a system they are working on to investigate a potential harm. Over the course of the interviews, a third window into ethical agency emerged, which we are calling "veiled agency."" (Griffins et al., 2023, p.5)

Regarding the technical feasibility of ethical demands, perceptions were split. The authors notice that a majority of the interviewed developers consider ethical principles can be coded into an algorithm but only mentioned this in relation to some principles like explainability.

Furthermore, they claim the importance of the "need to assign a metric or a number to the principle, and then iterate it." (Griffins et al., 2023, p.6)

About the authority to intervene, a majority does think they have authority. However, this authority "is tied to their positionality" (ibid), e.g., if they are junior or senior developers. Finally, there was a "veiled agency" aspect referring to the ethical agency "veiled" as technical agency:

> "Developers would sometimes list the myriad technical choices they make in the design, training, and deployment of automated systems. Occasionally, they would acknowledge they were also navigating ethical territories." (Griffins et al., 2023, p.6)

This ties back to their personal aims and perceptions. They might sense that there are ethical issues they should deal with, but they find it hard to identify these issues, know how to solve them or understand how they fit into their job. This confusion makes it difficult for them to step back and see the bigger picture —similar to the idea of looking up from the microscope discussed in Chapter 4. This difficulty arises not because they do not care about ethics, but because the developers' professional deformation (based on Polyakova's notion mentioned in section 4.3.2) limited their scope of analysis, for which interdisciplinary support is necessary. These insights into ethical agency are important because by acknowledging their role as ethical agents, developers strengthen their participation in active responsibility.

Along the same line, Griffin et al.'s research highlights the need for a more structured framework that can support developers in navigating ethical assessments, clarify the scope of their ethical responsibilities, and provide actionable guidance.

More specifically, the authors notice that developers:

> "[…] are grappling with morally troublesome gaps between who they believe themselves to be and what they are doing. […] Even if developers think of themselves as guided only by technical rules, this research reveals that personal morality still influences their sense of moral action and that they are engaged in ethical decision-making while they are developing automated systems." (Griffin et al., 2023, pp. 9-10)

Therefore, methodologies like the BNA can support AI developers in incorporating ethical considerations into their workflow, promoting a culture of reflection and teamwork, and enabling them to integrate their moral perspectives alongside the technical objectives of their projects. Thus, they can follow a sense of personal morality within a structured intervention to support them, and by doing so, they can be held accountable for their decision-making.

6.2.3 Responsible for past and future states of affairs.

I consider the discussion presented so far to be a strong grounding for forward- and backward-looking responsibilities to interact in the context of the application of the BNA.

By adopting an active responsibility stance, developers adopting the BNA can be held accountable for past actions in relation to the expectations of forward-looking responsibility, as noted in section 6.2.1.

Thus, regarding the ethical decision-making for addressing biases in AI development, developers can be held accountable for their choices, actions, and consequences when this corresponds. They would also be responsible for preventing harm they can foresee after applying the BNA. This active responsibility also allows them to develop a disposition to act responsibly.

As discussed in the pilot case study analysis, ethical decision-making becomes an inherent, rather than a burdensome aspect of the AI development cycle for developers. This implies acknowledging the constant presence of ethical choices in their practice; the ethical dimension is not an afterthought but a built-in feature of the development process.

In the pilot case study, I recognised that the development team faced constraints that were outside their immediate sphere of control. These constraints originated from various external factors and decisions that shaped the environment in which the developers operated. An understanding of these limitations was crucial.

For instance, the developers had to navigate the challenges posed by inherently biased gynaecological data sets, the prevalent professional biases within the medical field, and the legal restrictions dictated by Chile's transparency laws.

By acknowledging and analysing how these factors affected ethical considerations, the development team was able to pinpoint possible biases in their AI model and comprehend the broader implications. This awareness allowed them to adopt an active responsibility stance, managing expectations more effectively and facilitating transparent communication with other stakeholders regarding the potential impact of these biases on the AI system's performance and decision-making processes. They were able to recognise what was, as Tollon suggested, within their own control, therefore explicitly tracking potential risks that could depend, for example, on the deployment and use of the AI system by other stakeholders.

Hence, it is possible to delineate certain boundaries of this active responsibility in relation to applying the BNA. When the BNA is applied, we can expect developers to be responsible for the choices and foreseeable impacts that the approach allows them to recognise, i.e., to have a backward- and forward-looking responsibility about things that are under their control, such as addressing biases and their influence in the AI development process.

Accordingly, when posed with the challenge of determining negligence, for example, concerning the decisions made by the developer team, the BNA can be instrumental in clarifying who should be held responsible and why. While my proposal does not define how to establish responsibility or negligence, I suggest the transparency and introspection promoted by the BNA can contribute to this task.[17]

Overall, my claim is that the utilisation of the BNA as a methodological tool empowers developers to actively engage with the professional responsibility, we can attribute to them regarding the task of addressing biases. Thus, developers should recognise personal and systemic biases, the boundaries of their professional expertise, and the various elements that steer technical choices in the specific circumstances of their development context. By methodically delineating these aspects and their interrelations, the BNA enables developers to deepen their ethical decision-making.

With a more detailed understanding of the sociotechnical landscape given by interdisciplinary collaboration, developers are better equipped to foresee ethical pitfalls and address them proactively as part of their standard practice. This forward- looking approach goes beyond identifying immediate technical challenges; it involves identifying how AI systems and influencing biases interact with complex social dynamics and what ripple effects they might have.

[17] It is worth noticing, that AI developers operate within a complex ecosystem of many different stakeholders, such as their peers, project leaders, heads of organisations, users, policymakers, regulatory bodies, and the public at large. Each actor plays a significant role in ensuring AI is used and managed responsibly. Understanding the nuanced interactions between these roles and acknowledging the limits of what a single developer can influence is key to establishing a fair and effective framework for assigning responsibility in AI development. Such a framework would need to address both the responsibility for past actions and decisions (backward-looking) as well as the obligations towards future outcomes (forward-looking). My proposal merely recognises this broader context and offers a way to promote an active responsibility of developers to address biases. Nonetheless, this general consideration emphasises that developers are pivotal in the ethical construction of AI systems and that adopting a sociotechnical perspective also means acknowledging a notion of collective responsibility. This means all involved parties are accountable for their individual contributions to the ethical and societal impact of AI technology. Such a collective sense of responsibility could foster a more comprehensive and anticipatory approach to the ethical challenges in AI, promoting collaborative efforts to mitigate risks and enhance benefits for society.

Adopting the BNA, I argue, promotes an active responsibility that empowers development teams to create AI systems that are not only technically sound but also ethically robust. By continuously engaging with the ethical dimensions of their work, individual developers can ensure that their AI systems are aligned with societal values. As such, the bias network approach is not just a particular intervention for developing AI; it is a comprehensive framework that nurtures the ethical consciousness of AI developers.

6.3 Responsibilities from adopting the BNA.

The BNA considers the intricate interplay between technology and society, acknowledging that AI systems are not isolated entities but are deeply embedded within organisational and societal structures. By taking into account the social dynamics, power relations, and cultural contexts in which technology operates, a sociotechnical perspective, I have argued, allows for a more nuanced understanding of responsibility.

The approach encourages developers to engage in thorough contemplation about possible biases, significant influential factors, and their implications. Such a proactive approach establishes a framework where developers are accountable for making ethical choices right from the beginning of a project, necessitating a solid rationale for their design and developmental strategies, which can help explain the limitations of certain outcomes or recognise the potential risks. But there can also be other actors involved in the adoption of this approach.

6.3.1 Individuals.

Here I will comment on the responsibilities of developers adopting the BNA and prompters facilitating the approach.

6.3.1.1 Developers.

Developers have a set of responsibilities that come from adopting the approach.

The BNA requires them to actively engage with the ethical and societal aspects of AI. They are expected to consider the sociotechnical context as part of their developing process. However, because developers are not necessarily experts in societal or ethical aspects of AI development, we cannot make them responsible for not having expertise in this area. They can be responsible, however, for accounting for that sociotechnical context when adopting the BNA. When developers adopt the BNA, they have a responsibility to listen to the input from the prompter team and inform procedural choices with that input and the findings that come up mapping the network of biases.

Thus, developers are responsible for identifying and mapping out potential and existing biases, as well as influential bias-related factors within the development process. Mapping biases requires them to recognise when certain considerations might escape their professional knowledge and therefore consult other experts that might help evaluate the project. (This could happen organically with the interventions of the prompter team, but it could require specific input from other disciplinary experts, e.g., teachers, nurses, lawyers, etc.).

Developers are also responsible for upholding transparency standards. They should document their decision-making process to make it transparent and accessible. This refers to tracking decisions based on the influence or interaction of biases, and realisations regarding their professional limitations (e.g., the professional bias identified in the case study). Thus, AI developers must understand their own team's composition, including the range of expertise and prior experiences with high-impact social projects and interdisciplinary collaboration.

This self-assessment is crucial for identifying both strengths and potential weaknesses within the team, especially in ethical decision-making and addressing societal impacts. Once biases are mapped in their network for a specific project, developers are tasked with implementing interventions to address these biases, ensuring that AI systems function equitably and ethically. However, on certain occasions, they might be unable to tackle structural biases or other elements that are causing the problem (e.g., institutional bureaucracy or professional biases affecting data quality at the source, like in the case of gynaecology in the case study). These limitations need to be recognised, communicated, reported, and analysed to see how they might be affecting the AI lifecycle and development ecosystem. They have, therefore, a responsibility to inform these appropriately.

On certain occasions, it is the developers' responsibility to engage with various stakeholders, including the intended users of the technology and those affected by its deployment, to understand and incorporate diverse perspectives into the development process.

This was also a consideration raised in the pilot case study, where the development team saw a need to engage directly with the Ministry of Health professionals in charge of applying the new model, as this can provide further insight to prioritise implementation and deal with professional and cultural biases. Therefore, developers are also responsible for establishing a link with relevant stakeholders when necessary and possible.

Thus, the BNA places a forward-looking responsibility on developers to prevent undesirable effects of the technology they are creating. Furthermore, developers are expected to contribute to a transparent dialogue about the ethical implications of AI, making sure to report any bias-related issues they have encountered that raise ethical concerns.

This transparency is pivotal for fostering public trust and ensuring that AI development aligns with democratic values. It is worth emphasising, nonetheless, that if developers do not use the BNA, they are still responsible for considering the broader sociotechnical implications of their work. Even without the BNA, developers still have a responsibility to ensure their work is ethically sound and socially responsible.

The BNA is presented as a specific methodological approach that could help them address these dimensions more effectively to respond to their responsibilities about addressing bias. Therefore, even if they do not adopt the BNA, they remain responsible for the societal impacts of the AI systems they develop.

Here I just commented on specific aspects of the BNA for which they can be accountable.

6.3.1.2 Prompters.

Prompters, as described in Chapter 5, are professionals external to the developer team whose job is to guide the reflective process of the developers, ensuring they consider relevant questions about biases and bias-related factors.

Based on this consideration, prompters have a forward-looking responsibility to ensure they guide AI developers to ask relevant questions about biases and other contextual elements and factors related to biases, as well as make sure they integrate their reflections into the visualisation and documentation process. In this sense, it is essentially a responsibility to make sure they are “diligent prompters.”

In practice, this includes understanding the AI project in-depth, preparing appropriate prompts to facilitate reflective discussions, and guiding the developer team through a structured exploration of biases. They should be well-versed with the project’s scope, goals, and potential societal impacts.

Additionally, prompters are tasked with crafting questions that are open-ended, relevant, and sensitive to the developers’ disciplinary backgrounds, whilst being actively conscious of possible biases they could introduce into the analysis. Therefore, the prompters’ role is not to dictate solutions but to enable developers to see the broader implications of their work and encourage a multidisciplinary perspective, ensuring a comprehensive approach to ethical AI development.

But prompters can also have backward-looking responsibilities. Prompters might be negligent. If they are not diligent, they could be held accountable for failing to fulfil their roles. In certain cases, if this lack of due diligence results in a direct influence on developers’ decision-making which then has a negative impact on society, they can also be held accountable for this, if they were capable of performing well but failed to do so.

However, if prompters are not capable of fulfilling their role, because of a lack of resources, cooperation from the developers, or lack of skills, then they could be considered partially responsible or not responsible at all depending on the circumstances.

On certain occasions, it could be argued that prompters might have a responsibility that extends beyond facilitation; they may need to safeguard the process of implementing the BNA.

In cases where they observe the developer team disregarding critical advice or ethical guidelines to consider certain issues, as well as dismissing key information or not reporting found limitations, they may need to act.

This action can include reporting these issues to higher management, an ethics committee, or another relevant authority within the organisation in which the project is being developed. The aim is to ensure accountability and adherence to the ethical standards set forth for the project. The act of reporting is not necessarily accusatory, but a relevant measure to maintain the integrity of the development process and safeguard the responsible creation and implementation of AI technologies.

As regulatory landscapes evolve, especially with laws that focus on the pre-emptive management of risks associated with AI technologies, the role of prompters in managing biases becomes more critical, as they might be able to point out critical aspects necessary to be reported or fixed for a project to be approved or to avoid prosecution in the future.

Legislation that emphasises prevention, like the prospective personal data and AI regulations in Chile, implies that developer teams must be diligent in reporting project aims, assessing risks, and outlining mitigation strategies. In such a context, methodologies like the BNA become very valuable.

For organisations that may otherwise be inclined to prioritise profit over ethics, the "Evil Corp" type of institutions, adopting the BNA could be a strategic move to avoid legal repercussions.

6.3.2 Institutions.

In this section, I will discuss the responsibilities that developer organisations and associated institutions have when it comes to implementing the BNA.

6.3.2.1 Developer Companies.

When developer companies adopt the BNA as a standard practice, they are assuming a commitment to a set of forward and backward-looking responsibilities. Forward-looking responsibilities can include proactive measures such as continuously reviewing how the BNA is integrated into their AI development workflows.

For instance, they should ensure that their prompter teams are applying the BNA appropriately and that developer teams are following guidance and keeping the visualisation and documentation updated. Backwards-looking responsibilities could involve reflecting on and addressing any biases that have been identified in AI systems post-development, even with the BNA in place, as this could inform the prompter and developer teams for future interventions.

Ultimately, developer companies are responsible for shaping their organisational strategies for AI development around the lessons learned from the BNA. This means not just correcting past errors but also improving their development practices to prevent the recurrence of similar problems rooted in bias.

This is especially critical for companies whose primary business is the creation of AI technologies. They are expected to set industry standards in ethical AI development by actively seeking out potential biases and continuously evolving their methods to manage, mitigate and prevent them.

6.3.2.2 Educational Institutions.

Educational institutions, like universities and research centres, have a particular responsibility in shaping the future of AI development through their curriculum and research support. Incorporating the BNA into their academic programs can be an important step in preparing future generations of developers and professionals to competently address the sociotechnical challenges of addressing biases in AI development.

Furthermore, they should consider funding, resources, or offering other types of institutional support to research projects implementing the BNA, to make sure they have the necessary requirements to have a prompter team accompanying developers. For example, by having committees or protocols to oversee the implementation of the BNA as a requirement for AI research projects that have a direct societal impact.

6.3.2.3 Professional Institutions.

Professional institutions such as industry bodies or professional colleges, can also play a significant role in endorsing and facilitating the adoption of the BNA and other similar sociotechnical methodologies. They can include the BNA as part of their professional codes of ethics and regulations, as a way to promote ethical development.

Accordingly, these institutions could bear the responsibility of guiding their members in adopting the approach, by endorsing it as a standard practice and professional standard. They can also provide training for developers and possibly prompters, to learn more about sociotechnical approaches and the BNA. Or they could also offer network opportunities, bringing together professionals interested in applying a sociotechnical approach to their development processes, finding support, and exchanging ideas with others.

For example, a professional body could change its code of ethics to include specific sociotechnical references to address biases, highlighting the importance of methodologies like the BNA. Similarly, it could issue a requirement for its members to demonstrate competence in developing sociotechnical systems as part of their professional accreditation.

6.3.2.4 General institutional responsibilities.

Across institutions, there is an associated responsibility for them to foster a sociotechnical culture. The BNA is not just a technical checklist but an interdisciplinary endeavour that requires dialogue and reflection. Institutions should promote environments where AI developers, prompters, and stakeholders are part of the development process.

Unlike approaches focused on remediation, which often lead to quick technical fixes and temporary solutions, the BNA encourages developers to address biases systematically, preventing recurrent issues and understanding bias-related problems from a sociotechnical stance. Therefore, institutional responsibility also concerns supporting developers, prompters, and the organisation as a whole to adopt the BNA and assume the consequences if this implementation fails.

6.4 Conclusions.

Here I have presented some insights into the responsibilities associated with adopting the BNA. This approach embraces a sociotechnical perspective engaging with deeper societal implications, the discussion of the interaction of forward- and backward-looking responsibilities, as well as understanding technology within its social context.

By fostering an active responsibility through the implementation of the BNA, individuals and institutions are better equipped to create AI systems that support sociotechnical values. The proactive stance discussed here feeds on the interdisciplinary collaboration the BNA promotes, which enables developers to transcend their technical expertise and address the wider ethical implications of AI systems. Thus, the role of developers as ethical agents implies an active engagement in ethical decision-making to address biases.

The responsibilities of adopting the BNA are shared across the different actors involved in its adoption. Institutions have a pivotal role in supporting the implementation of the BNA. The specific responsibilities of developers and prompters are linked to their roles as decision-makers and facilitators respectively. These considerations foster a culture that values and integrates ethical considerations into AI development.

Furthermore, I have mentioned how there can be reasons for Evil Corp to adopt an approach like the BNA. Based on the analysis here, some of the reasons a developer or company otherwise not motivated to address ethical concerns could consider adopting the BNA.

First, there are strategic reasons as to why they could consider this. The BNA can be a sort of pre-emptive measure for legal repercussions. As regulation advances, more requirements are being set worldwide to increase transparency and involvement with ethical standards in AI development.

Therefore, an approach like the BNA can respond to those demands, offering interdisciplinary engagement to resolve and prevent bias-related issues as well as provide further insights about the justification for any decision-making by the developer team, fostering increased transparency and accountability. Thus, in the eyes of regulatory bodies, this could prevent sanctions related to biased or irresponsible practices.

Another potential reason that could convince Evil Corp is enhancing their customers' trust. The reputation of companies developing AI systems is increasingly questioned by avid users, foundations supporting the rights of consumers, or even activists concerned with ethical development standards. Adopting an approach like the BNA would mean an effective step towards committing the company to ethical standards and societal values, opening new markets and customer bases.

Lastly, one of the benefits of the BNA —already discussed in the thesis— is that by systematically identifying and mitigating biases, and considering related elements and factors, Evil Corp could reduce the risk of recurrent bias issues and prevent failures triggered by them. This is not only an ethical benefit in terms of responsible development, but also in making a better product, as applying the BNA makes the development of the system more likely to be reliable for the context in which it will be used. Thus, the BNA can be seen as a competitive advantage, making Evil Corp known for producing ethical AI solutions.

Final Conclusions

In this thesis, I have presented the Bias Network Approach (BNA) as a novel sociotechnical approach designed to aid AI developers in identifying and addressing biases more comprehensively. To achieve this, in the first half of the thesis, I provided a theoretical background to support the need for a sociotechnical approach, as well as some key distinctions that allowed me to argue in favour of a broader understanding of bias.

In Chapter 1, I criticised what I have called the problems of bias, i.e., technocentrism, conceptual ambiguity, and the isolationist approach —a problem that had not been previously discussed in the main literature. I also highlighted the limitations of these problems and how they can have a direct impact on how developers conceptualise bias. This helped me lay the groundwork for the argument that a more integrated sociotechnical approach is necessary for AI ethics in general, but specifically to discuss AI bias discussed in the next chapter.

In Chapter 2, I explained how understanding AI as a sociotechnical system and adopting a sociotechnical approach offers the possibility of having a broader context of analysis to deal with AI bias. I examined current critiques against AI ethics principles and guidelines, a theoretical proposal for redefining societal bias emphasising the importance of intersectional work, and a sociotechnical systematic approach to bias integrating feminist epistemology. From these, I gathered important insights to both recognise which sociotechnical elements were needed in my BNA proposal, as well as argue for a needed paradigm shift towards a sociotechnical approach for addressing the multifaceted nature of bias in AI development.

In Chapter 3, I introduced the most common categorisations for AI bias and presented two views on it, the negative (defining bias as bad) and the alternative view (claiming not all biases are bad, i.e., some can be positive). I concluded that the alternative views contribute to conceptual ambiguities by misconceptualising bias. I highlighted the importance of a balanced understanding of bias and considered the possibility of “repurposing” biases in a positive way, but without altering their negative denotation.

I also suggested a working definition of bias based on it being a deviation of error and a systematic tendency that must be understood by AI developers as an interconnected element within the AI ecosystem.

In Chapter 4, I introduced the BNA and detailed the methodology used to analyse the case study implemented to test it. I demonstrated the BNA’s potential to enhance the developers’ ethical awareness to address biases, and therefore improve their decision-making through core findings that included: the important role of discussing professional biases in AI, the relevance of having a multifactorial approach that allowed developers to visualise, discuss, and map biases and other relevant factors, and the contribution of the BNA to increase transparency and communicate the developers’ ethical decisions.

In Chapter 5, I provided initial guidance for the application of the BNA in AI development projects, detailing stages of implementation and highlighting the flexibility and adaptability of the approach. I also argued for some advantages of adopting the BNA, such as including the avoidance of superficial technical fixes by promoting a comprehensive view of biases; supporting the documentation of ethical decision-making, to help ensure rigorous ethical standards; and facilitating transparent and accessible communication, involving different stakeholders and post-deployment monitoring.

In Chapter 6, I described how responsibility as a moral obligation was related to the forward- and backward-looking responsibilities that come from adopting the BNA. More specifically, I also commented on how the ethical agency of developers related to their engagement in active responsibility and I discussed the specific responsibilities attributed to the main actors involved in adopting the BNA, noticing how they all share a responsibility to ensure the BNA is implemented appropriately by fostering a sociotechnical culture.

At the end, I considered some of the reasons that the likes of Evil Corp could have to adopt the BNA, considering the benefits of a sociotechnical approach that prevents and mitigates bias-related risks as well as increase transparency and accountability in line with evolving regulatory requirements and societal expectations for ethical AI development.

In summary, in this thesis, I presented a comprehensive argument for the adoption of the Bias Network Approach (BNA) as a critical tool to help AI developers address biases within AI development. Through a detailed examination of different challenges surrounding AI bias, I argued that the BNA is part of a necessary and fundamental sociotechnical shift, where solutions to bias-related problems are understood within a broader context integrating societal, ethical, and technical considerations. Here, I have also shown how the approach can promote a culture of ethical awareness and proactive responsibility among AI developers.

I consider, therefore, the BNA proposal to be more than just an example of how a sociotechnical approach could be translated into an intervention; it is a call to action for developers to become aware of and oversee the ethical implications of their work. The BNA offers a comprehensive approach that respects the importance of technical expertise of AI developers, and at the same time strengthens their ethical commitments by collaborating in an interdisciplinary setting.

Finally, I would like to refer to a recurrent implication I discussed throughout the thesis, in reference to a "technocentric mindset" adopted at large by tech companies and institutions, and by AI developers.

The assertion that AI engineers and the companies employing them exhibit a technocentric and isolationist tendency to deal with biases does not imply a moral deficit per se. As identified in the pilot case study, we need to acknowledge the existence of systemic and cultural tendencies within the field. The "microscopic vision" referenced in Chapter 5 underscores the predisposition of engineering disciplines to narrow their focus on specific technical problems, often neglecting the broader ethical and societal dimensions of their work. However, there is a growing recognition within the AI community that these limitations hinder the development of equitable, reliable systems, that is, how ethical consideration improve the precision and quality of AI models, thereby opening the door to more holistic approaches such as the Bias Network Approach (BNA).

AI engineers and companies are increasingly aware of the shortcomings of technocentric solutions.

Purely technical fixes, such as fairness metrics or dataset balancing, have demonstrated limited efficacy in addressing the structural and systemic dimensions of bias. The insufficiency of these approaches has led to public scrutiny and reputational risks for companies, creating an imperative to adopt methodologies that go beyond surface-level adjustments. The BNA provides a framework for identifying the root causes of bias within and across the levels of AI development, from data collection to system implementation, making it an attractive solution for addressing these deeper challenges.

Despite its demanding nature, the BNA offers substantial pragmatic benefits. By identifying and addressing biases at multiple levels, the approach not only improves model robustness but also enhances transparency and accountability—qualities increasingly demanded by regulators and stakeholders. The iterative mapping of biases within the BNA fosters risk management, allowing teams to identify and address potential pitfalls before they result in significant errors or societal harms. As such, the BNA aligns with both the operational goals of AI companies and their ethical obligations.

External pressures further incentivise the adoption of the BNA. Regulatory frameworks governing AI systems, particularly in sensitive sectors such as healthcare, finance, and public policy, are becoming stricter, necessitating comprehensive and transparent approaches to bias mitigation. At the same time, market differentiation increasingly hinges on demonstrating ethical AI practices. Companies that proactively adopt interventions such as the BNA can position themselves as leaders in ethical innovation, attracting socially conscious consumers and securing their reputation in a competitive market. Additionally, the BNA aligns with the goals of academic and institutional partnerships, fostering collaborations that can provide both funding opportunities and reputational benefits.

Moreover, cultural shifts within the engineering and AI communities suggest a growing openness to interdisciplinary and inclusive methodologies. The traditional technocentric mindset is not immutable; engineers and developers are increasingly recognising the importance of engaging with social sciences, ethics, and stakeholder perspectives. The collaborative, interdisciplinary nature of the BNA serves as a bridge between these fields, fostering an organisational culture that prioritises ethical and systemic considerations alongside technical precision.

In conclusion, the endorsement of the Bias Network Approach by AI engineers and companies, even those perceived as “Evil Corp” is not only plausible but strategically advantageous. By addressing the limitations of the technocentric mindset and aligning with regulatory, market, and cultural shifts, the BNA provides a pathway for advancing ethical and responsible AI development and the complex challenges of dealing with biases in AI.

Afterword: Future Work and a "Holistic Network Approach"

Here I would like to touch briefly on future work, highlighting the importance of continuous improvement in methodologies for ethical AI development. One immediate avenue for future work involves gathering more case studies to test the BNA. This would provide further insights to improve its implementation.

Another, perhaps less evident avenue, is using the basis of the BNA to explore other ways in which the concept of a "network approach" could help AI developers.

As this thesis unfolded, what I initially constructed as an idea to improve how AI developers think about bias issues in their professional role, showed promise in another more general achievement: making ethical thinking more accessible to AI developers. After the intervention, the developers participating in the case study manifested a change in how they were thinking about their projects, and not only about bias. By using the BNA to think about biases and their interconnections in the AI pipeline, they were drawn to consider a broader context, which required them to think about ethics more generally.

This is why I see the potential to develop an analogous network approach but with a broader spectrum, a "holistic network approach".

Derived from the Greek word *holos* meaning "whole", the term holistic denotes a perspective that considers an entire system or entity, instead of just individual parts. Therefore, it relates to a sense of interconnected or interdependent aspects that constitute that system. Accordingly, sociotechnical approaches to AI can be identified as holistic because they are used to analyse AI's technical and societal aspects, grounding them in a contextual assessment.

Different authors have characterised their take on sociotechnical perspectives for AI systems as a holistic account. Akbarighatar et al. (2023) for example, adopt a sociotechnical perspective to develop a framework that "provides a holistic approach for addressing both instrumental and humanistic objectives of AI development" (Akbarighatar et al., 2023, p.10). The authors claim that one of the reasons they adopt a sociotechnical perspective is that it provides “a unique approach to understanding holistically by the capabilities required for harnessing the power of AI while minimizing the risks for societies and individuals” (ibid, p.11).

Ehsan and Riedl (2020), explain how a sociotechnical approach helps them achieve a holistic understanding of explainable AI systems, accounting for the socially situated nature whilst requiring technical and social insights. For the authors, a sociotechnical approach helps them:

> “[…] critically reflect or contemplate on implicit or unconscious values embedded in computing practices […] Such contemplation—or reflection— can bring unconscious or implicit values and practices to conscious awareness, making them actionable. As a result, we can design and evaluate technology in a way that is sensitive to the values of both designers and stakeholders” (Ehsan & Riedl, 2020, p.450).

Hence, when I refer to thinking holistically in a sociotechnical AI context, I mean considering different elements and how they fit together when designing AI systems—i.e., including a broader context that goes beyond technical considerations. Consider that adopting a holistic network approach is like building a puzzle; rather than concentrating on individual pieces, we must pay attention to how they connect, i.e., how they fit to build a full picture. Therefore, a different version of the network approach could employ a holistic viewpoint to map the interrelations between diverse ethical concerns in AI development.

Going deeper into the puzzle metaphor, the fundamental meaning of holistic lies in recognising how the pieces of the puzzle shift as we build the picture. The evolution or adaptation of ethical considerations can also be understood in a metaphorical sense.

When we build a puzzle, the pieces do not magically change shape, the puzzle “changes” as it makes sense because we have put pieces together that create the necessary context for the individual pieces to acquire meaning, i.e., forming an image. Mapping different ethical concerns could offer a similar benefit, that is, help developers see the whole picture.

The holistic network approach could be designed to help the developer team visualise how different ethical concerns interact with one another, to document their decision-making processes and to justify their reflective practices.

For example, the holistic version of the network approach could help developers identify and weigh in potential conflicts or trade-offs regarding issues about privacy. For instance, when prioritising user privacy, there may be limitations in the amount of data available for enhancing algorithms, creating a challenge that affects both privacy and performance considerations affecting users—a sociotechnical concern.

Consequently, the holistic network approach could allow developers to get valuable insights about ethically significant factors impacting how the developer team frames their development process. Therefore, the holistic network approach could also aid in establishing a foundational grasp of the ethical landscape for the AI project and its implementation context.

Nonetheless, as I previously discussed in the main work of this thesis, it would be unrealistic to assume that these developers would spontaneously shift their focus from technical to broader ethical considerations (looking up from the microscope) unless they have undergone appropriate training.

This is precisely why external interventions are conducted by a prompting team. These interventions can raise awareness within the developer team, enabling them to recognise the ethical landscape.

But to understand this proposal for future work better, I will explain in more depth why I consider the holistic approach a potentially valuable method.

Imagine you are part of a developer team developing an AI system designed to aid medical practitioners with treatments and diagnoses. Similar to the developers in the case study, your team's strengths are on the technical side, you have professionals with engineering and medical knowledge, and the team knows how to make the AI system suitable for the task.

But, like the researchers in the pilot, you wish to know how to approach ethically relevant questions that might not be evident to you and that are related to how well the model can help patients and doctors, beyond optimisation goals. Hence, your goal is to have further insight into ethical distinctions that can help you and your team be able to justify and explain decisions, that although at times are technical, still have an important societal dimension.[18]

The foundational premise of the holistic approach then, much like the BNA did with biases, is that ethical considerations and other relevant factors are not considered in isolation. The procedural considerations would require the team to address and think about privacy, fairness, transparency, and even sustainability issues from the get-go, thus the holistic approach can help them map how these concerns can interact throughout the entire development process.

In practice, this translates to the developer team identifying how the AI system could interact with the users, in this case, the medical practitioners, and those being affected by the system, i.e., the patients.[19] This would allow them to contextualise the AI system design and identify relevant concerns related to their development.

To do so, the concept of patient care is a central aspect of implementing AI systems in healthcare, and part of that patient care experience is rooted in the level of trust patients have in their practitioners as well as the AI system being used by them.

Hence, the developer team will need to grasp how to build trust in this context (or at least promote it). Here, I will show how starting from trust as a focal point, the team could map the interaction of other ethical elements that could be interacting and,

[18] Notice that I emphasise the willingness of the developer team. I assume that the proposal of these methodological approaches will be adopted by research teams or institutions that care to improve their processes and gain further insight into the ethical aspects of their development procedures, as mentioned in the main work of this thesis.

[19] Here I just mention the most directly affected stakeholder. However, it is possible to recognise other indirect or less obvious, such as the healthcare system. In the AI ethics literature, the identification and typification of stakeholders have been widely discussed in the last few years. Some references are the role of different stakeholders in relation to explainable AI (McDermid et al., 2021); identification of passive and representative stakeholder roles in AI projects (Miller, 2022); building responsible AI systems and their potential stakeholders (Deshpande & Sharp, 2022).

therefore, influencing the realisation of this trust —keeping in mind this is just a hypothetical for future work and, therefore, this example would need further considerations to become an alternative network approach.

Trust has been recognised in the literature as a mechanism that can influence and shape how medical practitioners and patients adopt and benefit from AI in healthcare contexts (Asan et al., 2020). Therefore, trust is closely related to our epistemic accessibility, how much we know and how certain we are of this knowledge—in this case, concerning what the AI system does and what the user (the medical practitioner) gets from that use.

Hence, patients will require a certain degree of transparency about the quality of input data, for example, or the basic functioning of the AI system, to know what it does and how it does it, to then understand how the medical practitioner weighs that into her decision-making.

Following that epistemic requirement, there are at least two basic components of trust AI developers should care to promote when designing their system: (i) technical reliability (e.g., the precision of the AI systems), and (ii) the interpersonal aspects of trust (e.g., experiential memories and possible biases). It is also important to balance trust expectations, as blind trust is not necessarily the best outcome for proper human-AI interaction.

A better expectation could be what Asan et al. (2020) call "optimal trust", to reflect the necessary level of scepticism for both AI and humans, as they both are capable of erring. This translates into a requirement for AI developers, to incorporate mechanisms that can support and maintain that level of optimal trust, which is based on the capabilities and limitations of the AI system and should be aligned with the intended use the medical practitioner should give to the system.

Now, there are different ways in which developers can promote this optimal trust. For example, to establish a secure use of patient data for model training, there is a demand for transparency and privacy concerns.

In this case, legitimising informed consent must be required. Furthermore, security and privacy guidelines can connect with transparency requirements. As discussed in the analysis of the pilot case study in Chapter 4, there can be important differences in the type of data, its quality, and the biases this data can contain (e.g., professional biases affecting gynaecological data).

Ensuring transparency about these data limitations is related to fairness and bias because these aspects are relevant to the AI system's performance and representativity. So, if there is limited data on gastroenterology diseases in autistic or elderly patients, for example, then the medical practitioner must be informed on how poorly the model could work on that part of the population, so that the practitioner can decide how to use the system or if using it would be appropriate, based on her professional expertise—i.e., decide if the AI system can be trusted for the task.

If developers have these limitations in mind, by applying a holistic network, they can think about and map the different alternatives available to correct potential issues of trust from the technical side, but also document and communicate those technical limitations and the societal dimension they influence as well. Such considerations can be complemented with explainable AI mechanisms, for example, that can promote informed judgements by users and patients.

Moreover, when the decisions made by the developer team are transparent, this can also facilitate identifying responsibilities. If a biased decision occurs, transparency allows for tracing the responsibility back to the developers, data sources, algorithmic choices, or even external actors and material limitations (as seen in the case study in Chapter 4), thus fostering accountability.

What the network approaches could provide is a way to systematically incorporate these context-specific elements into the development process, and into how developers are thinking about these ethical concerns within a sociotechnical network. By systematically observing how ethical concerns interact in a given context, the developer team can gain a comprehensive understanding to help them navigate potential ethical dilemmas and sociotechnical considerations.

Consider, for example, the efforts to achieve an optimal trust discussed above, these could be mapped as shown in Figure 11. A much more detailed mapping would require a robust analysis, but this shows a rudimentary image of the possible connections.

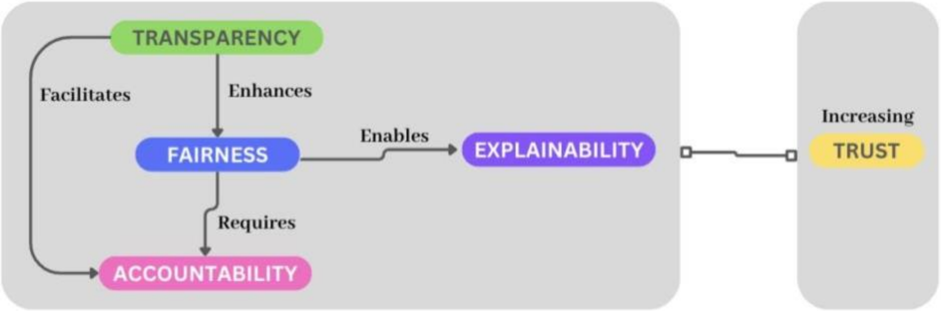


Figure 10:Rudimentary holistic network diagram for the AI system focused on medical diagnoses and treatment recommendations for which increasing trust is a set goal.

This method of illustrating the interaction of different ethical concerns could help map how they can complement and reinforce each other. It reveals that these different ethical aspects create the essential conditions for each other's functioning. The holistic network approach could uncover the inherent connections between them, emphasising a critical aspect of adopting a sociotechnical perspective: that issues and elements related to transparency, fairness, or explainability should not be understood in isolation, but as part of a network of influences.

In simple words, the idea of a "holistic network approach" can be translated to how we could map "everything" relevant to sociotechnical ethical decision-making in AI development, from technical limitations and opportunities to societal biases, institutional limitations, the needs of different stakeholders, potential foreseeable risks like physical or psychological harm, etc. Adapting this, however, might be challenging, of course, as this would be more complex than mapping just biases and bias-related concerns, however, the fundamental benefits of "the network approach" could still hold.

The core idea is to promote interdisciplinary dialogue and interventions, avoid isolationist approaches to thinking about sociotechnical issues in AI, and map this into a network of influences, where each node represents a facet of the complex interplay between technology and society. This visualisation can offer a more holistic understanding of the ethical landscape AI developers must face, identifying issues as they arise, but also considering how different elements may interact in unexpected ways, promoting a forward-looking responsibility.

References

Adasme, S., Arriagada, G., Lopez, C., & Pertuze, J. (Forthcoming). APEC Digital Economic Steering Group (DESG), Comparative study on best practices to detect and avoid harmful biases in Artificial Intelligence systems. Asia- Pacific Economic Cooperation (APEC).

Akbarighatar, P., Pappas, I., & Vassilakopoulou, P. (2023). A sociotechnical perspective for responsible AI maturity models: Findings from a mixed- method literature review. International Journal of Information Management Data Insights, 3(2), 100193. https://doi.org/10.1016/j.jjimei.2023.100193

Akter, S., Dwivedi, Y. K., Biswas, K., Michael, K., Bandara, R. J., & Sajib, S. (2021). Addressing Algorithmic Bias in AI-Driven Customer Management. Journal of Global Information Management (JGIM), 29(6), 1–27. https://doi.org/10.4018/JGIM.20211101.oa3

Akter, S., McCarthy, G., Sajib, S., Michael, K., Dwivedi, Y. K., D'Ambra, J., & Shen, K. N. (2021). Algorithmic bias in data-driven innovation in the age of AI. International Journal of Information Management, 60, 102387. https://doi.org/10.1016/j.ijinfomgt.2021.102387

APEC Digital Economy Steering Group. (2023). Best Practices to Detect and Avoid Harmful Biases in Artificial Intelligence Systems (APEC Project: DESG 05 2021A). APEC. https://www.apec.org/docs/default-source/publications/2023/9/223_desg_best-practices-to-detect-and-avoid-harmful-biases-in-artificial-intelligence-systems.pdf

Arriagada-Bruneau, G., López, C. & Davidoff, A. A Bias Network Approach (BNA) to Encourage Ethical Reflection Among AI Developers. Sci Eng Ethics 31, 1 (2025). https://doi.org/10.1007/s11948-024-00526-9

Antony, L. M. (2016). Bias: Friend or Foe?: Reflections on Saulish Skepticism. In Implicit Bias and Philosophy, Volume 1. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198713241.003.0007

Asan, O., Bayrak, A. E., & Choudhury, A. (2020). Artificial Intelligence and Human Trust in Healthcare: Focus on Clinicians. Journal of Medical Internet Research, 22(6), e15154. https://doi.org/10.2196/15154

Báez, P., Bravo-Marquez, F., Dunstan, J., Rojas, M., & Villena, F. (2022). Automatic Extraction of Nested Entities in Clinical Referrals in Spanish. ACM Transactions on Computing for Healthcare, 3(3), 28:1-28:22. https://doi.org/10.1145/3498324

Baker, R. S., & Hawn, A. (2021). Algorithmic Bias in Education. EdArXiv. https://doi.org/10.35542/osf.io/pbmvz

Barad, K. (2007). Meeting the Universe Halfway: Quantum Physics and the Entanglement of Matter and Meaning. Duke University Press.

Barocas, S., & Selbst, A. D. (2016). Big Data's Disparate Impact. California Law Review, 104(671). https://doi.org/10.2139/ssrn.2477899

Benk, M., Tolmeijer, S., von Wangenheim, F., & Ferrario, A. (2022). The Value of Measuring Trust in AI - A Socio-Technical System Perspective (arXiv:2204.13480). arXiv. https://doi.org/10.48550/arXiv.2204.13480

Bietti, E. (2020). From ethics washing to ethics bashing: A view on tech ethics from within moral philosophy. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 210–219. https://doi.org/10.1145/3351095.3372860

Bijker, W. E., & Law, J. (1992). Shaping technology/building society: Studies in sociotechnical change. The MIT Press. https://hdl.handle.net/2027/heb01128.0001.001

Boddington, P. (2017). Towards a Code of Ethics for Artificial Intelligence (1st ed.). Springer Cham. https://doi.org/10.1007/978-3-319-60648-4

Brandon, J. (2021). Using unethical data to build a more ethical world. AI and Ethics, 1(2), 101–108. https://doi.org/10.1007/s43681-020-00006-3

Braun, M., Hummel, P., Beck, S., & Dabrock, P. (2021). Primer on an ethics of AI- based decision support systems in the clinic. Journal of Medical Ethics, 47(12), e3–e3. https://doi.org/10.1136/medethics-2019-105860

Broussard, M. (2018). Artificial Unintelligence: How Computers Misunderstand the World. MIT Press.

Broussard, M. (2019, June 17). Letting Go of Technochauvinism. Public Books. https://www.publicbooks.org/letting-go-of-technochauvinism/

Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. Proceedings of the 1st Conference on Fairness, Accountability and Transparency, 77–91. https://proceedings.mlr.press/v81/buolamwini18a.html

Char, D. S., Abràmoff, M. D., & Feudtner, C. (2020). Identifying Ethical Considerations for Machine Learning Healthcare Applications. The American Journal of Bioethics: AJOB, 20(11), 7–17. https://doi.org/10.1080/15265161.2020.1819469

Chen, Z. (2023). Ethics and discrimination in artificial intelligence-enabled recruitment practices. Humanities and Social Sciences Communications, 10(1), Article 1. https://doi.org/10.1057/s41599-023-02079-x

Clark, A. (2024). The Experience Machine: How Our Minds Predict and Shape Reality. Random House.

Clark, A. A nice surprise? Predictive processing and the active pursuit of novelty. *Phenom Cogn Sci* 17, 521–534 (2018). https://doi.org/10.1007/s11097-017-9525-z

Clark, A. (2016). *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*. Oxford University Press.

Clegg, C. W. (2000). Sociotechnical principles for system design. Applied Ergonomics, 31(5), 463–477. https://doi.org/10.1016/s0003-6870(00)00009-0

Coeckelbergh, M. (2020). Artificial Intelligence, Responsibility Attribution, and a Relational Justification of Explainability. Science and Engineering Ethics, 26(4), 2051–2068. https://doi.org/10.1007/s11948-019-00146-8

Coglianese, C., & Lai, A. (2022). Algorithm vs. Algorithm (SSRN Scholarly Paper 4026207). https://papers.ssrn.com/abstract=4026207

Cramer, H., Garcia-Gathright, J., Springer, A., & Reddy, S. (2018). Assessing and addressing algorithmic bias in practice. Interactions, 25(6), 58–63. https://doi.org/10.1145/3278156

Danks, D., & London, A. J. (2017). Algorithmic Bias in Autonomous Systems. Proceedings of the 26th International Joint Conference on Artificial Intelligence, 4691–4697. http://dl.acm.org/citation.cfm?id=3171837.3171944

Davis, M. (1998). Thinking Like an Engineer: Studies in the Ethics of a Profession. Oxford University Press.

Davis, M. C., Challenger, R., Jayewardene, D. N. W., & Clegg, C. W. (2014). Advancing socio-technical systems thinking: A call for bravery. Applied Ergonomics, 45(2), 171–180. https://doi.org/10.1016/j.apergo.2013.02.009

Deshpande, A., & Sharp, H. (2022). Responsible AI Systems: Who are the Stakeholders? AIES '22: Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society, 227–236. https://oro.open.ac.uk/84505/

Dobbe, R., Dean, S., Gilbert, T., & Kohli, N. (2018). A Broader View on Bias in Automated Decision-Making: Reflecting on Epistemology and Dynamics. In arXiv e-prints. https://ui.adsabs.harvard.edu/abs/2018arXiv180700553D

Douglas, H. E. (2009). Science, Policy, and the Value-Free Ideal. University of Pittsburgh Press. https://doi.org/10.2307/j.ctt6wrc78

Draude, C., Klumbyte, G., Lücking, P., & Treusch, P. (2019). Situated algorithms: A sociotechnical systemic approach to bias. Online Information Review, 44(2), 325–342. https://doi.org/10.1108/OIR-10-2018-0332

Ehsan, U., & Riedl, M. O. (2020). Human-Centered Explainable AI: Towards a Reflective Sociotechnical Approach. In C. Stephanidis, M. Kurosu, H. Degen, & L. Reinerman-Jones (Eds.), HCI International 2020—Late Breaking Papers: Multimodality and Intelligence (pp. 449–466). Springer International Publishing. https://doi.org/10.1007/978-3-030-60117-1_33

Estay, R., Cuadrado, C., Crispi, F., González, F., Alvarado, F., & Cabrera, N. (2017). Desde el conflicto de listas de espera, hacia el fortalecimiento de los prestadores públicos de salud: Una propuesta para Chile. Cuadernos Médico Sociales, 57(1), Article 1.

Fabi, S., & Hagendorff, T. (2022). Why we need biased AI -- How including cognitive and ethical machine biases can enhance AI systems. arXiv:2203.09911 [Cs]. http://arxiv.org/abs/2203.09911

Fazelpour, S., & Danks, D. (2021). Algorithmic bias: Senses, sources, solutions. Philosophy Compass, 16(8), e12760. https://doi.org/10.1111/phc3.12760

Ferrara, E. (2023). Fairness And Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, And Mitigation Strategies (arXiv:2304.07683). arXiv. https://doi.org/10.48550/arXiv.2304.07683

Feuerriegel, S., Dolata, M., & Schwabe, G. (2020). Fair AI. Business & Information Systems Engineering, 62(4), 379–384. https://doi.org/10.1007/s12599-020-00650-3

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M. (2020). Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI (SSRN Scholarly Paper 3518482). Social Science Research Network. https://doi.org/10.2139/ssrn.3518482

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2018, March 23). Datasheets for Datasets. arXiv.Org. https://arxiv.org/abs/1803.09010v8

Gichoya, J. W., Meltzer, C., Newsome, J., Correa, R., Trivedi, H., Banerjee, I., Davis, M., & Celi, L. A. (2022). Ethical Considerations of Artificial Intelligence Applications in Healthcare. In C. N. De Cecco, M. van Assen, & T. Leiner (Eds.), Artificial Intelligence in Cardiothoracic Imaging (pp. 561– 565). Springer International Publishing. https://doi.org/10.1007/978-3-030- 92087-6_52

Gichoya, J. W., Thomas, K., Celi, L. A., Safdar, N., Banerjee, I., Banja, J. D., Seyyed-Kalantari, L., Trivedi, H., & Purkayastha, S. (2023). AI pitfalls and what not to do: Mitigating bias in AI. The British Journal of Radiology, 96(1150), 20230023. https://doi.org/10.1259/bjr.20230023

Green, B., & Viljoen, S. (2020). Algorithmic realism: Expanding the boundaries of algorithmic thought. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 19–31. https://doi.org/10.1145/3351095.3372840

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. Psychological Review, 102(1), 4–27. https://doi.org/10.1037/0033-295X.102.1.4

Griffin, T. A., Green, B. P., & Welie, J. V. M. (2023). The ethical agency of AI developers. AI and Ethics. https://doi.org/10.1007/s43681-022-00256-3

Hagendorff, T. (2020) The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds & Machines* 30, 99–120. https://doi.org/10.1007/s11023-020-09517-8

Haslanger, S. (2015). Distinguished Lecture: Social structure, narrative and explanation. Canadian Journal of Philosophy, 45(1), 1–15. https://doi.org/10.1080/00455091.2015.1019176

Haraway, D. (1988). Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective. Feminist Studies, 14(3), 575–599. https://doi.org/10.2307/3178066

Harding, S. (1992). Rethinking Standpoint Epistemology: What Is "Strong Objectivity"? In Feminist Epistemologies. Routledge.

Harding, S. (2015). Objectivity and Diversity: Another Logic of Scientific Research. University of Chicago Press.

Hellström, T., Dignum, V., & Bensch, S. (2020). Bias in Machine Learning What is it Good (and Bad) for? arXiv:2004.00686 [Cs]. http://arxiv.org/abs/2004.00686

Hooker, S. (2021). Moving beyond "algorithmic bias is a data problem". Patterns, 2(4), 100241. https://doi.org/10.1016/j.patter.2021.100241

Jiang, H., & Nachum, O. (2019). Identifying and Correcting Label Bias in Machine Learning. CoRR, abs/1901.04966. http://arxiv.org/abs/1901.04966

Johnson, G. M. (2020). Algorithmic Bias: On the Implicit Biases of Social Technology. Synthese, 198(10), 9941–9961. https://doi.org/10.1007/s11229- 020-02696-y

Kelly, T. (2022). Bias: A Philosophical Study. Oxford University Press.

Kizilcec, R. F., & Lee, H. (2022). Algorithmic fairness in education. In The Ethics of Artificial Intelligence in Education. Routledge.

Krafft, T. D., Zweig, K. A., & König, P. D. (2022). How to regulate algorithmic decision-making: A framework of regulatory requirements for different applications. Regulation & Governance, 16(1), 119–136. https://doi.org/10.1111/rego.12369

Kroll, J. A., Huey, J., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G., & Yu, H. (2017). Accountable Algorithms. University of Pennsylvania Law Review, 165(3), 633.

Lecaros, C., Dunstan, J., Villena, F., Ashcroft, D. M., Parisi, R., Griffiths, C. E. M., Härtel, S., Maul, J. T., & De la Cruz, C. (2021). The incidence of psoriasis in Chile: An analysis of the National Waiting List Repository. Clinical and Experimental Dermatology, 46(7), 1262–1269. https://doi.org/10.1111/ced.14713

Liao, S., Zhang, R., Poblete, B., & Murdock, V. (2023). Bias Invariant Approaches for Improving Word Embedding Fairness. Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, 1400–1410. https://doi.org/10.1145/3583780.3614792

Lin, Y. (2012). Transdisciplinarity and Digital Humanities: Lessons Learned from Developing Text-Mining Tools for Textual Analysis. In D. M. Berry (Ed.), Understanding Digital Humanities (pp. 295–314). Palgrave Macmillan UK. https://doi.org/10.1057/9780230371934_16

Lloyd, K. (2018). Bias Amplification in Artificial Intelligence Systems. arXiv:1809.07842

Loi, M., Ferrario, A., & Viganò, E. (2021). Transparency as design publicity: Explaining and justifying inscrutable algorithms. Ethics and Information Technology, 23(3), 253–263. https://doi.org/10.1007/s10676-020-09564-w

Longino, H. E. (1993). Feminist Standpoint Theory and the Problems of Knowledge. Signs, 19(1), 201–212.

McDermid, J. A., Jia, Y., Porter, Z., & Habli, I. (2021). Artificial intelligence explainability: The technical and ethical dimensions. Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences, 379(2207), 20200363. https://doi.org/10.1098/rsta.2020.0363

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2019). A Survey on Bias and Fairness in Machine Learning. USC, Information Sciences Institute.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A Survey on Bias and Fairness in Machine Learning. ACM Computing Surveys, 54(6), 115:1-115:35. https://doi.org/10.1145/3457607

Miller, G. J. (2022). Stakeholder roles in artificial intelligence projects. Project Leadership and Society, 3, 100068. https://doi.org/10.1016/j.plas.2022.100068

Mitchell, S., Potash, E., Barocas, S., D'Amour, A., & Lum, K. (2021). Algorithmic Fairness: Choices, Assumptions, and Definitions. Annual Review of Statistics and Its Application, 8(1), 141–163. https://doi.org/10.1146/annurev-statistics-042720-125902

Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. Nature Machine Intelligence, 1(11), 501–507. https://doi.org/10.1038/s42256-019-0114-4

Morley, J., Elhalal, A., Garcia, F., Kinsey, L., Mökander, J., & Floridi, L. (2021). Ethics as a service: A pragmatic operationalisation of AI Ethics (SSRN Scholarly Paper 3784238). https://doi.org/10.2139/ssrn.3784238

Morozov, E. (2014). To Save Everything, Click Here. Technology, Solutionism, and the Urge to Fix Problems that Don't Exist. Penguin Random House.

Moser, A., & Korstjens, I. (2017). Series: Practical guidance to qualitative research. Part 1: Introduction. The European Journal of General Practice, 23(1), 271– 273. https://doi.org/10.1080/13814788.2017.1375093

Munn, L. (2023). The uselessness of AI ethics. AI and Ethics, 3(3), 869–877. https://doi.org/10.1007/s43681-022-00209-w

Nachtwey, O., & Seidl, T. (2023). The Solutionist Ethic and the Spirit of Digital Capitalism. Theory, Culture & Society, 02632764231196829. https://doi.org/10.1177/02632764231196829

Niehaus, F., & Wiesche, M. (2021). A Socio-Technical Perspective on Organizational Interaction with AI: A Literature Review. ECIS 2021 Research Papers. https://aisel.aisnet.org/ecis2021_rp/156

Ntoutsi, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdl, W., Vidal, M.-E., Ruggieri, S., Turini, F., Papadopoulos, S., Krasanakis, E., Kompatsiaris, I., Kinder-Kurlanda, K., Wagner, C., Karimi, F., Fernandez, M., Alani, H., Berendt, B., Kruegel, T., Heinze, C., Staab, S. (2020). Bias in data-driven artificial intelligence systems—An introductory survey. WIREs Data Mining and Knowledge Discovery, 10(3), e1356. https://doi.org/10.1002/widm.1356

Olteanu, A., Castillo, C., Diaz, F., & Kıcıman, E. (2019). Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries. Frontiers in Big Data, 2, 13. https://doi.org/10.3389/fdata.2019.00013

Parikh, R. B., Teeple, S., & Navathe, A. S. (2019). Addressing Bias in Artificial Intelligence in Health Care. JAMA, 322(24), 2377–2378. https://doi.org/10.1001/jama.2019.18058

Paul, R. (2022). Can critical policy studies outsmart AI? Research agenda on artificial intelligence technologies and public policy. Critical Policy Studies, 16(4), 497–509. https://doi.org/10.1080/19460171.2022.2123018

Paullada, A., Raji, I. D., Bender, E. M., Denton, E., & Hanna, A. (2021). Data and its (dis)contents: A survey of dataset development and use in machine learning research. Patterns, 2(11), 100336. https://doi.org/10.1016/j.patter.2021.100336

Peeters, M. M. M., van Diggelen, J., van den Bosch, K., Bronkhorst, A., Neerincx,M. A., Schraagen, J. M., & Raaijmakers, S. (2021). Hybrid collective intelligence in a human–AI society. AI & SOCIETY, 36(1), 217–238. https://doi.org/10.1007/s00146-020-01005-y

Polyakova, O. (2014). The Structure of Professional Deformation. Procedia - Social and Behavioral Sciences, 146, 420–425. https://doi.org/10.1016/j.sbspro.2014.08.148

Pot, M., Kieusseyan, N., & Prainsack, B. (2021). Not all biases are bad: Equitable and inequitable biases in machine learning and radiology. Insights into Imaging, 12(1), 13. https://doi.org/10.1186/s13244-020-00955-7

Rajpurkar, P., Chen, E., Banerjee, O., & Topol, E. J. (2022). AI in health and medicine. Nature Medicine, 28(1), Article 1. https://doi.org/10.1038/s41591- 021-01614-0

Releséguier, A., & Rodrigues, R. (2020). AI ethics should not remain toothless! A call to bring back the teeth of ethics. Big Data & Society, 7(2). https://doi.org/10.1177/2053951720942541

Richardson, B., & Gilbert, J. E. (2021). A Framework for Fairness: A Systematic Review of Existing Fair AI Solutions. https://doi.org/10.48550/arXiv.2112.05700

Rocher, L., Hendrickx, J. M., & de Montjoye, Y.-A. (2019). Estimating the success of re-identifications in incomplete datasets using generative models. Nature Communications, 10(1). https://doi.org/10.1038/s41467-019-10933-3

Roselli, D., Matthews, J., & Talagala, N. (2019). Managing Bias in AI. Companion Proceedings of The 2019 World Wide Web Conference, 539–544. https://doi.org/10.1145/3308560.3317590

Rovatsos, M., Mittelstadt, B., & Koene, A. (2019). Landscape Summary: Bias in Algorithmic Decision-Making: What is bias in algorithmic decision-making, how can we identify it, and how can we mitigate it? UK Government. https://www.gov.uk/government/publications/landscape-summaries commissionedby-the-centre-for-data-ethics-and-innovation

Sangokoya, D. (2020). Algorithmic accountability – Applying the concept to different country contexts. World Wide Web Foundation and Data Pop Alliance. https://datapopalliance.org/publications/algorithmic-accountability- applying-the-concept-to-different-country-contexts/

Santoni de Sio, F., & Mecacci, G. (2021). Four Responsibility Gaps with Artificial Intelligence: Why they Matter and How to Address them. Philosophy & Technology, 34(4), 1057–1084. https://doi.org/10.1007/s13347-021-00450-x

Saul, J. (2013). Scepticism and Implicit Bias. Disputatio, 5(37), 243–263.

Schwartz, R., Vassilev, A., Greene, K., Perine, L., Burt, A., & Hall, P. (2022).Towards a Standard for Identifying and Managing Bias in Artificial Intelligence. National Institute of Standards and Technology, U.S. Department of commerce. https://doi.org/10.6028/NIST.SP.1270

Smith, G., & Rustagi, I. (2020). Mitigating Bias in Artificial Intelligence. An Equity Fluent Leadership Playbook. https://haas.berkeley.edu/equity/industry/playbooks/mitigating-bias-in-ai/

Soleimani, M., Intezari, A., Taskin, N., & Pauleen, D. (2021). Cognitive biases in developing biased Artificial Intelligence recruitment system. Hawaii International Conference on System Sciences 2021 (HICSS-54). https://aisel.aisnet.org/hicss-54/ks/big_data_analytics/3

Song, J. Y., Pycha, A., & Culleton, T. (2022). Interactions between voice-activated AI assistants and human speakers and their implications for second-language acquisition. Frontiers in Communication, 7. https://www.frontiersin.org/articles/10.3389/fcomm.2022.995475

Srinivasan, R., & Chander, A. (2021). Biases in AI systems. Communications of the ACM, 64(8), 44–49. https://doi.org/10.1145/3464903

Striphas, T. (2015). Algorithmic culture. European Journal of Cultural Studies, 18(4–5), 395–412. https://doi.org/10.1177/1367549415577392

Suresh, H., & Guttag, J. (2021). A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. Equity and Access in Algorithms, Mechanisms, and Optimization, 1–9. https://doi.org/10.1145/3465416.3483305

Talbert, M. (2023). Moral Responsibility. In E. N. Zalta & U. Nodelman (Eds.), The Stanford Encyclopedia of Philosophy (Fall 2023). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2023/entries/moral-responsibility/

Tollon, F. (2022). Is AI a Problem for Forward Looking Moral Responsibility? The Problem Followed by a Solution. In E. Jembere, A. J. Gerber, S. Viriri, & A. Pillay (Eds.), Artificial Intelligence Research (pp. 307–318). Springer International Publishing.

United States Home Office. (2016). Big data: A report on algorithmic systems, opportunity, and civil rights. The White House. https://purl.fdlp.gov/GPO/gpo90618

van de Poel, I. (2011). The Relation Between Forward-Looking and Backward- Looking Responsibility. In N. A. Vincent, I. van de Poel, & J. van den Hoven (Eds.), Moral Responsibility: Beyond Free Will and Determinism (pp. 37– 52). Springer Netherlands. https://doi.org/10.1007/978-94-007-1878-4_3

van de Poel, I. (2020). Embedding Values in Artificial Intelligence (AI) Systems. Minds and Machines, 30(3), 385–409. https://doi.org/10.1007/s11023-02009537-4

West, S. M., Whittaker, M., & Crawford, K. (2019). Discriminating Systems: Gender, Race and Power in AI. AI Now Institute. https://ainowinstitute.org/publication/discriminating-systems-gender-race- and-power-in-ai-2

Whittlestone, J., Nyrup, R., Alexandrova, A., Dihal, K., & Cave, S. (2019). Ethical and societal implications of algorithms, data, and artificial intelligence: A roadmap for research. Leverhulme Centre for the Future of Intelligence, University of Cambridge. https://www.nuffieldfoundation.org/sites/default/files/files/Ethical-and- Societal-Implications-of-Data-and-AI-report-Nuffield-Foundat.pdf

Wilson, M., & Frank, R. (2023). Inductive Bias Is in the Eye of the Beholder. In D. Hupkes, V. Dankers, K. Batsuren, K. Sinha, A. Kazemnejad, C.

Christodoulopoulos, R. Cotterell, & E. Bruni (Eds.), Proceedings of the 1st GenBench Workshop on (Benchmarking) Generalisation in NLP (pp. 152– 162). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.genbench-1.12

Zajko, M. (2022). Artificial intelligence, algorithms, and social inequality: Sociological contributions to contemporary debates. Sociology Compass, 16(3), e12962. https://doi.org/10.1111/soc4.12962

Završnik, A. (2021). Algorithmic justice: Algorithms and big data in criminal justice settings. European Journal of Criminology, 18(5), 623–642. https://doi.org/10.1177/1477370819876762

Zhai, X., & Krajcik, J. (2023). Pseudo AI Bias (SSRN Scholarly Paper 4368917). https://doi.org/10.2139/ssrn.4368917

Zhou, Y., Kantarcioglu, M., & Clifton, C. (2023). On Improving Fairness of AI Models with Synthetic Minority Oversampling Techniques. In Proceedings of the 2023 SIAM International Conference on Data Mining (SDM) (pp. 874– 882). https://doi.org/10.1137/1.9781611977653.ch98

Zwakman, D. S., Pal, D., & Arpnikanondt, C. (2021). Usability Evaluation of Artificial Intelligence-Based Voice Assistants: The Case of Amazon Alexa. SN Computer Science, 2(1), 28. https://doi.org/10.1007/s42979-020-00424-4

Annex.

Table 1. Literature review to build the standard pipeline for AI development.

Authors and date	Title	Pipeline stages
Akter et al. (2021)	Algorithmic bias in data-driven innovation in the age of AI	1. Data product conceptualization 2. Data acquisition and refinement 3. Data storage and retrieval 4. Data product distribution 5. Data product presentation 6. Market feedback
Baker & Hawn ac (2021)	Algorithmic bias in education	1. Data generation 2. Task definition 3. Data measurement 4. Model learning 5. Evaluation and post-processing 6. Model Deployment 7. Feedback and stakeholders
Barocas & Selbst (2016)	Big data's disparate impact	1. Defining target variable 2. Data training 3. Feature selection
Char et al. (2020)	Identifying ethical considerations for machine learning healthcare applications	1. Conception 2. Development 3. Calibration 4. Inspection 5. Initial implementation 6. Evaluation
Cramer et al. (2018)	Assessing and addressing algorithmic bias in practice	1. Input data 2. Algorithm and team decisions 3. Results
Danks & London (2017)	Algorithmic bias in autonomous systems	1. Training data 2. Algorithm design 3. Implementation and interpretation

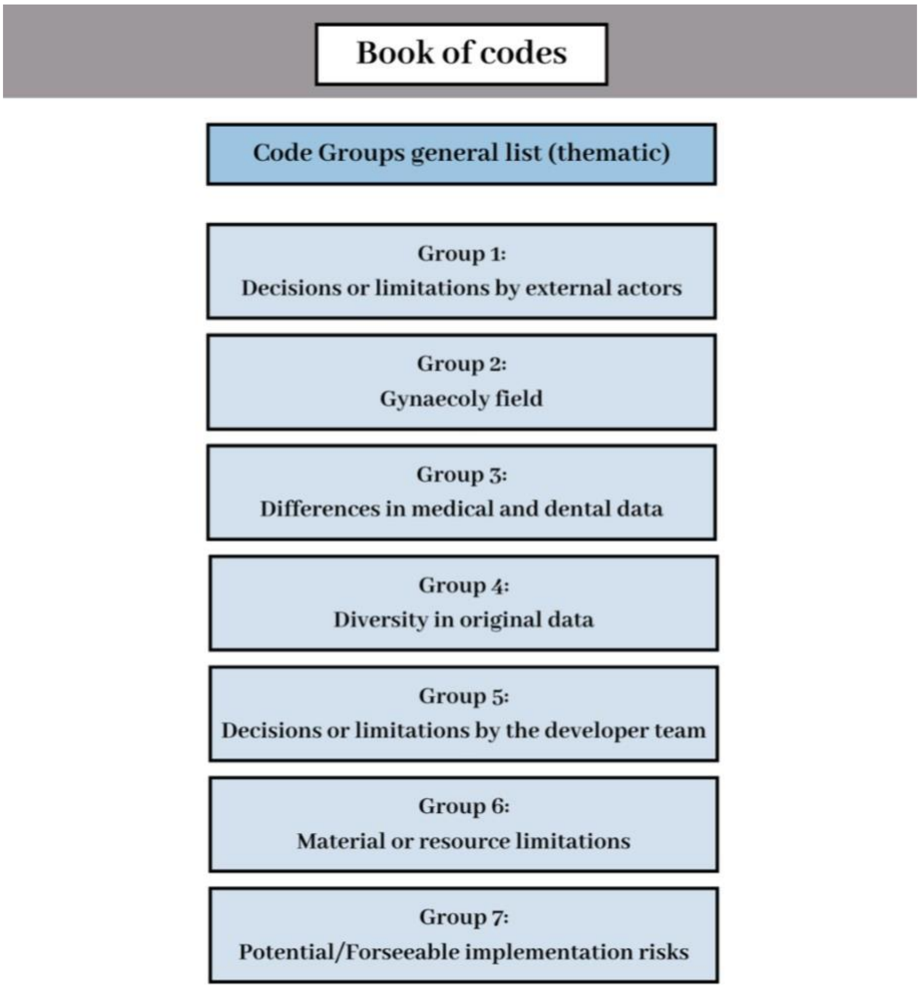
Dobbe et al. (2018)	A broader view on bias in automated decision-making: reflecting on epistemology and dynamics	1. Design 2. Development 3. Implementation
Draude et al. (2019)	Situated algorithms: a socio-technical systemic approach to bias	1. Data stage 2. Algorithmic design 3. Implementation
United States Home Office (2016)	Big data: a report on algorithmic systems, opportunity, and civil rights	1. Input stage 2. Design of algorithmic systems and machine learning
Fazelpour & Danks (2021)	Algorithmic bias: senses, sources, solutions	1. Problem specificationTa 2. Data collection and pre-processing 3. Modelling and validation 4. Deployment
Feuerriegel et al. (2020)	Fair AI: challenges and opportunities	1. Data 2. Modelling 3. Inadequate implementation 4. Model learning: 5. Evaluation and post-processing 6. Model Deployment 7. Feedback from stakeholders
Johnson (2020)	Algorithmic bias: on the implicit biases of social technology	No explicit pipeline Mentions data collection, labeling, and model design stages
Kizilcec and Lee (2022)	Algorithmic fairness in education. In the ethics of artificial intelligence in education	1. Problem definition and data 2. Model learning 3. Action (output and implementation)
Mehrabi et al. (2021)	A survey on bias and fairness in machine learning	1. Data 2. Algorithm 3. User interaction
Mitchell et al. (2021)	Algorithmic fairness: choices, assumptions, and definitions	1.Problem definition 2. Data

		3. Model design 4. Evaluation
Olteanu et al. (2019)	Social data: biases, methodological pitfalls, and ethical boundaries	1. Research design 2. Data collecting 3. Processing 4. Analysis 5. Evaluation
Parikh et al. (2019)	Addressing bias in artificial intelligence in health care	1. Data collection 2. Model design and implementation 3. Interpretation
Paullada et al. (2021)	Data and its (dis)contents: a survey of dataset development and use in machine learning research	1. Data collection 2. Data annotation 3. Documentation 4. Benchmarking 5. Data reuse
Richardson & Gilbert (2021)	A framework for fairness: a systematic review of existing fair AI solutions	1. Data collection 2. Data processing 3. Implementation and interpretation
Roselli et al. (2019)	Managing bias in AI	1. Goal definition 2. Data stage 3. Implementation
Rovatsos et al. (2019)	Landscape summary: bias in algorithmic decision-making	1. Input stage 2. Algorithm stage 3. Implementation and evaluation 4. Interpretation
Smith & Rustagi (2020)	Mitigating bias in artificial intelligence. an equity fluent leadership playbook	1. Data collection/ labelling 2. Algorithmic design and evaluation 3. Implementation and interpretation
Suresh & Guttag (2021)	A framework for understanding sources of harm throughout the machine learning life cycle	1. Data collection 2. Data preparation 3. Model development

		4. Model evaluation 5. Model post-processing 6. Model deployment
Sangokoya (2020)	Algorithmic accountability – Applying the concept to different country contexts	1. Input data 2. Processing and weighting of data 3. Implementation 4. Feedback
Srinivasan & Chander (2021)	Biases in AI systems	1. Data collection, labelling and pre-processing 2. Problem formulation 3. Algorithm and analysis 4. Testing and validation
Završnik (2021)	Algorithmic justice: Algorithms and big data in criminal justice settings	1. Database building 2. Algorithm design 3. Implementation 4. Interpretation 5. Feedback

Book of codes

Book of codes to make thematic group connections between type of problem or bias origin, divided into 7 distinctive groups identified in the waiting list project case study.



Book of codes

Code Groups general list (thematic)

Group 1:
Decisions or limitations by external actors

Group 2:
Gynaecoly field

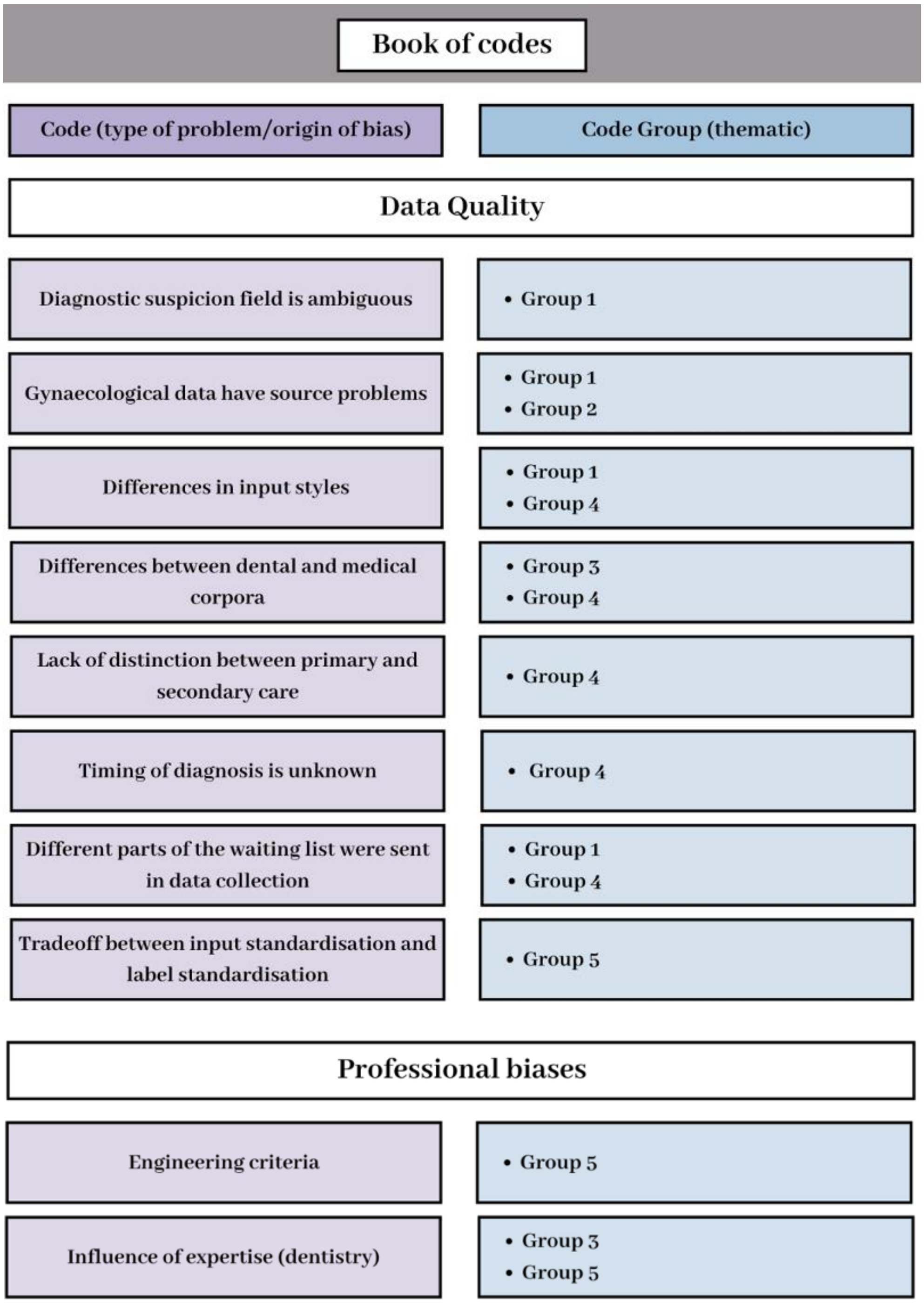
Group 3:
Differences in medical and dental data

Group 4:
Diversity in original data

Group 5:
Decisions or limitations by the developer team

Group 6:
Material or resource limitations

Group 7:
Potential/Forseeable implementation risks



Book of codes

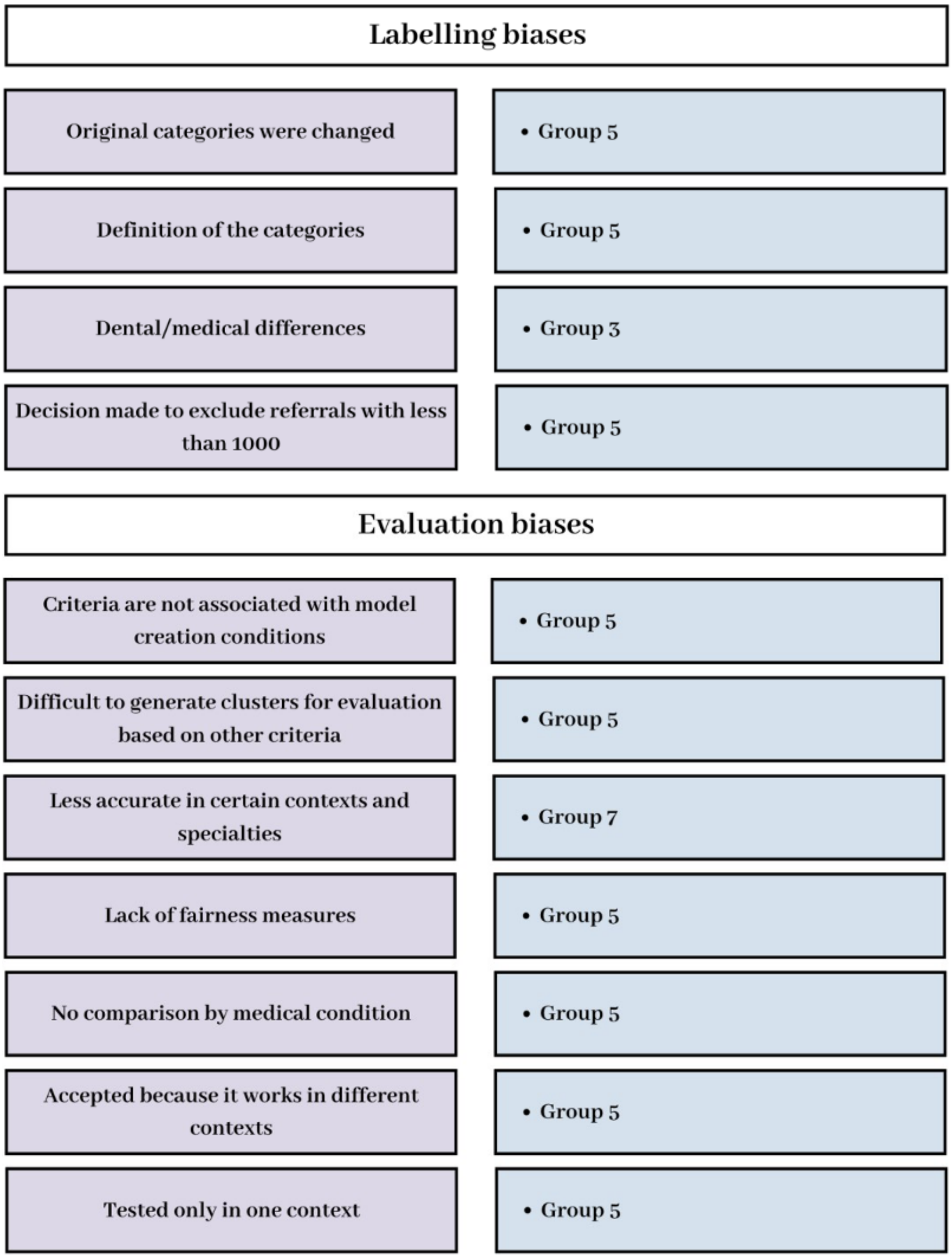
Code (type of problem/origin of bias)	Code Group (thematic)
Data Quality	
Diagnostic suspicion field is ambiguous	• Group 1
Gynaecological data have source problems	• Group 1 • Group 2
Differences in input styles	• Group 1 • Group 4
Differences between dental and medical corpora	• Group 3 • Group 4
Lack of distinction between primary and secondary care	• Group 4
Timing of diagnosis is unknown	• Group 4
Different parts of the waiting list were sent in data collection	• Group 1 • Group 4
Tradeoff between input standardisation and label standardisation	• Group 5
Professional biases	
Engineering criteria	• Group 5
Influence of expertise (dentistry)	• Group 3 • Group 5

Data collection

Missing data	• Group 1
Lack of reliability in data generation protocols in public hospitals	• Group 1
Difficulties to obtain data	• Group 1 • Group 6
Midwives generating data (over doctors)	• Group 1 • Group 2
Inconsistencies in data collection for gynaecological data	• Group 1 • Group 2

Data representativeness

Certain specialties generate more data than others	• Group 1
Certain specialties overrepresented due to data selection	• Group 1
Certain services provide more information than others	• Group 1
Data altered according to allocated public health resources	• Group 1 • Group 6
Decision made to achieve parity between dental and medical data	• Group 3 • Group 5



Labelling biases

Original categories were changed	• Group 5
Definition of the categories	• Group 5
Dental/medical differences	• Group 3
Decision made to exclude referrals with less than 1000	• Group 5

Evaluation biases

Criteria are not associated with model creation conditions	• Group 5
Difficult to generate clusters for evaluation based on other criteria	• Group 5
Less accurate in certain contexts and specialties	• Group 7
Lack of fairness measures	• Group 5
No comparison by medical condition	• Group 5
Accepted because it works in different contexts	• Group 5
Tested only in one context	• Group 5

Implementation biases

Difference in performance for medical and dental contexts	• Group 3
No consideration for positive/negative impacts	• Group 5
Unknown criteria of implementation by future actors	• Group 5 • Group 1
Considerations are put on quantity not gravity of conditions	• Group 5
Applications in contexts different on the one intended by the research team	• Group 1 • Group 7

Societal biases

Socio-ecnomic inequalities	• Group 1 • Group 6
Lack of human and material resources for the developer team	• Group 6
Institutional interests of the data holder	• Group 1 • Group 6
Technical deficiencies of the data holder	• Group 1 • Group 6
Inaccurate gynaecological terms	• Group 2

Interpretation biases

Not all public administrative workers will know how to use it / apply it	• Group 1
Not all healthcare workers will know how to use it	• Group 1

Model design

Criteria are defined based on the data	• Group 5
Differences between data types are not considered	• Group 5
Specialties are not differentiated	• Group 5