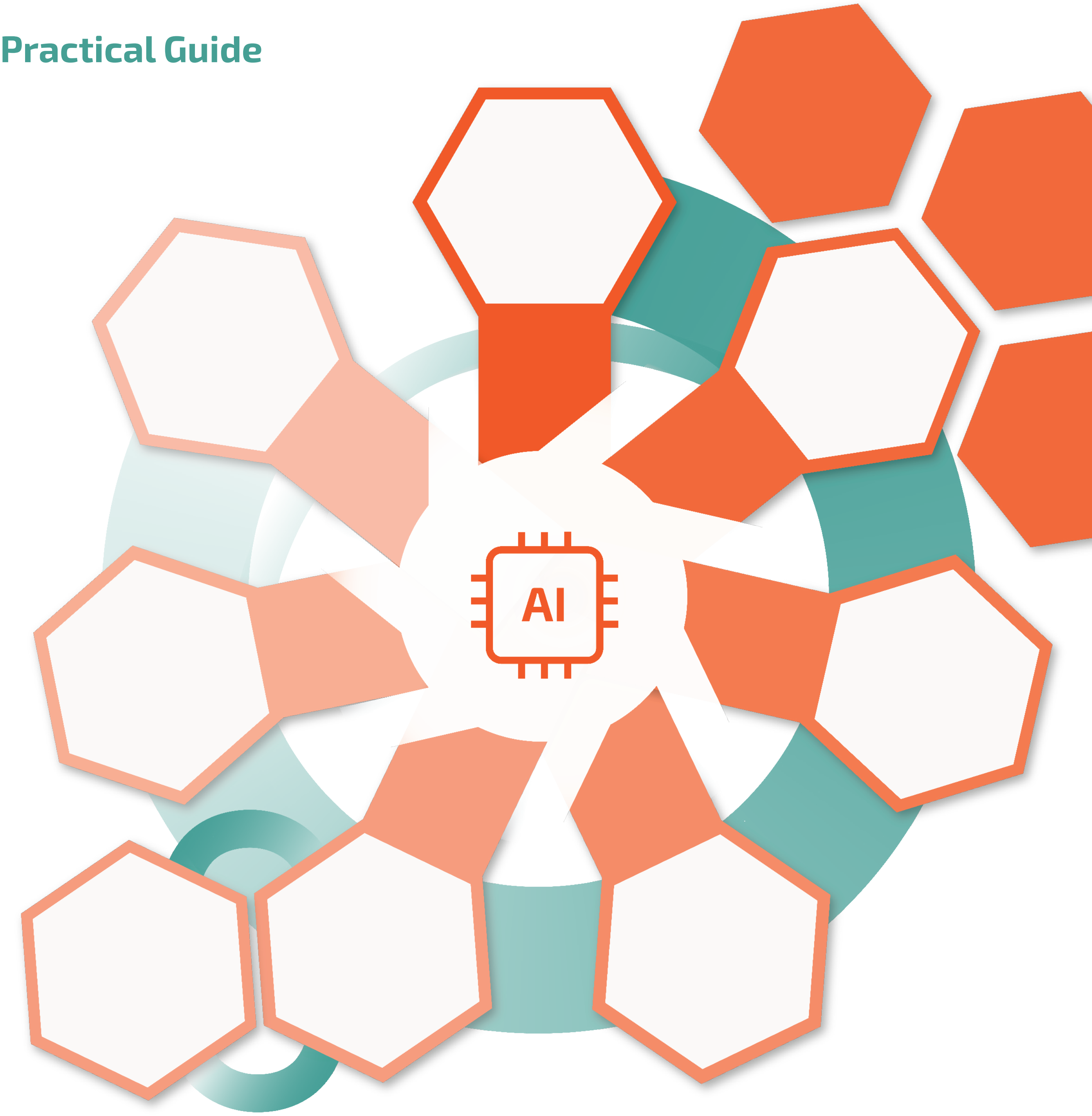
From Inception to Retirement: Addressing Bias Throughout the Lifecycle of AI Systems

A Practical Guide

Rhite
Leading the way to Trustworthy AI



Published by

Rhite
in collaboration with
Radboud Universiteit

Authors
Suzanne Snoek
Isabel Barberá
September 5, 2024

WE STAND BY RESPONSIBLE INNOVATION

At Rhite, we foster a collaborative environment that thrives on knowledge exchange and pioneering research. We strongly advocate for the responsible development of new technologies, dedicating significant resources to exploring how to make Trustworthy AI technically achievable.

WHY *THIS* PAPER?

As Artificial Intelligence (AI) systems become increasingly integrated into organizational decision-making processes, we have observed that organizations across both the public and private sectors—ranging from small startups to large enterprises—struggle to understand and address the biases inherent in these systems. Whether small businesses or large corporations or institutions, bias identification and mitigation remain a persistent challenge.

We initiated this research to address the following key challenges:

- **Organizational Lack of Knowledge**
  Many organizations lack the knowledge and confidence to identify and address biases in their AI systems. There is often uncertainty about the extent of bias present and hesitation in finding methods to mitigate potential issues.
- **Gaps in Sociotechnical Expertise**
  Developers and other stakeholders frequently lack the necessary expertise to recognize and address biases that stem from the intricate interaction between technical systems and social factors. This knowledge gap makes it difficult for them to understand what to look for and how to apply effective mitigation strategies that account for both technical solutions and the broader social context in which the AI operates.
- **Lack of Comprehensive Guidance**
  While some work has been done to identify biases in specific stages of the AI lifecycle, there is a need for a comprehensive resource that provides guidance on the different types of bias and mitigation measures throughout the entire lifecycle of an AI system. This study is the first one to offer such guidance, taking into account the eight stages of the AI lifecycle model as outlined in ISO/IEC 22989.

This research aims to fill these gaps by offering clear and actionable insights to help organizations and stakeholders better understand, identify, and address biases in AI systems ensuring more fair and equitable outcomes.

ABOUT US

Rhite
Leading the way to Trustworthy AI

Rhite is a consultancy and research firm spezialized in Trustworthy AI, with a focus on risks, governance, and compliance. Combining technical and legal expertise, Rhite uses a holistic, risk-based approach to ensure that AI technologies meet regulatory and technical standards managing risks and upholding human values.

Our expertise

We offer a unique blend of technical know-how and legal expertise in AI.
Our advisors stand out for their multidisciplinary approach to Trustworthy AI.

WHAT WE DO

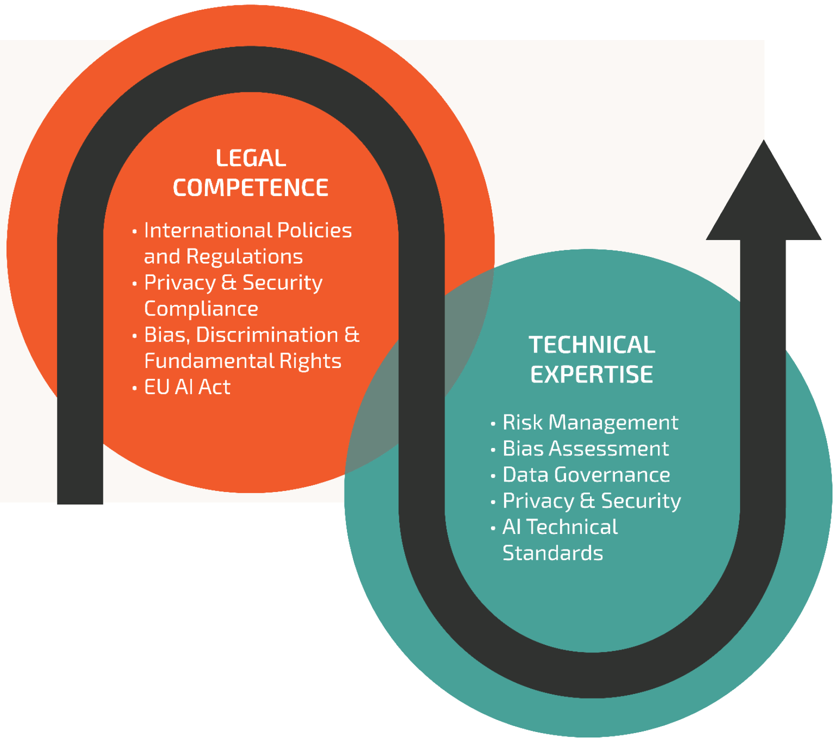
- Legal and technical consultancy on AI;
- Guidance to comply with the requirements of the EU AI Act;
- Auditing of algorithms and AI systems;
- Privacy, Security, safety and fundamental rights Impact assessments of AI solutions;
- Bespoke trainings on AI Risk Management;
- Implementation of Responsible AI programs.

HOW WE DO IT

RHITE is an acronym representing the principles we believe should underpin the design, development, and use of AI:

- **Responsible**
- **Humane**
- **Ingenious**
- **Transparent**
- **Empathic**

A holistic approach towards Trustworthy AI



LEGAL COMPETENCE

- International Policies and Regulations
- Privacy & Security Compliance
- Bias, Discrimination & Fundamental Rights
- EU AI Act

TECHNICAL EXPERTISE

- Risk Management
- Bias Assessment
- Data Governance
- Privacy & Security
- AI Technical Standards

Our founders

Isabel Barberá
Co-founder | AI advisor & Privacy Engineer

With a multidisciplinary background in privacy and security, engineering, AI, law and ethics, she guides organisations in the design and implementation of responsible digital solutions. She is an advocate of Trustworthy AI by design and passionate about the protection of human rights.

Martijn Korse
Co-founder | Privacy & Security Engineer

Martijn has a long career in the field of software engineering, DevSecOps and cybersecurity. Besides that, he also has a background in psychology and philosophy. Like Isabel, Martijn has a passion for privacy and security by design and he is also a strong advocate of responsible human-centered design.

Learn more about
us on our website!
www.rhite.tech

Table of Contents

Executive Summary

AI systems, while powerful and transformative, are susceptible to biases that can lead to unfair outcomes and discrimination. These biases can stem from multiple sources, including the data used to train models, the algorithms themselves, and even the decisions made by human stakeholders during the AI lifecycle. Unaddressed, these biases can perpetuate existing inequalities and create new ones, undermining trust in AI technologies and leading to significant societal harm.

This research provides a comprehensive guide to the different types of bias that can arise during each phase of the lifecycle of an AI system. It also offers practical recommendations for addressing these biases, ensuring that AI systems are developed, deployed, and managed in a fair and equitable manner.

Introduction

With the rapid advancement and increasing prominence of AI, it is crucial to recognize that, while AI holds the promise of numerous opportunities, it also carries significant risks. AI systems can be susceptible to a range of threats, including those related to security and safety. Moreover, they have the potential to contribute to and exacerbate environmental and human rights issues. Among these risks, the presence of bias in AI systems is particularly concerning.

Biases are deeply embedded in human nature and societal structures, making their existence in society inevitable. While some biases may be neutral or contextually appropriate, the presence of harmful biases within AI systems can lead to unfair treatment or discrimination against individuals, groups, and, if pervasive, can contribute to broader societal inequalities. There have been numerous instances where the application of AI systems has led to unjust outcomes or even outright discrimination. A prominent example is the Dutch Child Benefit Scandal, where the Dutch Tax and Customs Administration used algorithms as part of a larger system that disproportionately targeted low- and middle-income families applying for childcare benefits. Factors such as "foreign-sounding names" and "dual nationality" contributed to the algorithms' unfair targeting, leading to racial profiling, false accusations of fraud, and severe financial penalties for the affected families (Amnesty, 2021). While this scandal involved a complex interplay of factors beyond just a biased AI system, it underscores the profound impact that biased AI can have, highlighting the urgent need to address these issues as AI continues to evolve and integrate into more aspects of our daily lives.

SCOPE AND PURPOSE

The primary purpose of this document is to help AI developers, practitioners, and other stakeholders recognize the more common types of bias that can arise throughout the lifecycle of an AI system. By providing awareness and practical recommendations, this document aims to equip stakeholders with the knowledge needed to identify, address, and mitigate bias at every stage of AI development and deployment.

This document offers a comprehensive overview of the AI lifecycle, focusing on the critical phases where bias is most likely to emerge. While bias is often linked to issues of discrimination and fairness, this document will also clarify the distinctions between these concepts, emphasizing their unique roles and implications within AI systems. It is important to note that this document serves as an introductory guide rather than an exhaustive resource. While it provides valuable insights into various types of bias and offers general recommendations for mitigation, it does not explore every bias in depth or cover all possible strategies. To support further learning and action, references to additional information sources will be provided. Ultimately, this document is designed to foster a more informed and proactive approach among those involved in AI development and governance, ensuring that AI systems are designed and implemented with fairness and equity in mind.

BIAS

The concept of “bias” within the context of AI is complex and subject to varied interpretations, making it difficult to define explicitly. In general, bias can be understood as a “systematic difference in treatment of certain objects, people, or groups in comparison to others,” where “treatment” encompasses a wide range of actions, including perception, observation, representation, prediction, or decision-making. Essentially, bias reflects a systematic and disproportionate tendency towards a particular outcome or group (“ISO/IEC TR 24027:2021,” 2021).

Not all biases are inherently negative; some may be necessary for the functioning of an AI system. However, unintentional and unwanted biases can lead to unfair results, which can undermine the system’s fairness and equity. In the AI domain, the term “algorithmic bias” is often used to specifically refer to biases present within algorithms themselves (Kordzadeh & Ghasemaghaei, 2021).

Bias in AI can be understood through two primary lenses: technical and social. The technical framing of bias views it as a statistical phenomenon, one that can be addressed through data improvement and refined algorithm design. In contrast, the social framing extends beyond mere statistics, considering bias within the broader historical and political contexts. This perspective emphasizes the need to address bias through structural changes and a deeper understanding of the social power dynamics that influence AI systems (Ulnicane & Aden, 2023).

A significant challenge in minimizing unwanted bias lies in the fact that individuals are often unaware of their own biases, which may inadvertently influence AI systems. This lack of awareness makes it difficult to recognize and address biases during the development of AI systems. Therefore, it is crucial for developers and stakeholders to become vigilant about identifying potential biases, asking the right questions, and analyzing the possible consequences of these biases throughout the lifecycle of an AI algorithm.

DIFFERENT TYPES OF BIAS IN AI

Biases in AI systems can emerge when there is a disconnect between how reality is represented in data and how we perceive or idealize the world. Figure 1 illustrates the relationship between the two primary categories of bias: societal biases and statistical biases (Mitchell et al., 2021).

Societal Bias refers to the ingrained prejudices, stereotypes, or inclinations that are embedded within a culture or society. These biases influence perceptions, judgments, and behaviors towards certain groups or individuals (Schwartz et al., 2022). Societal biases can be either positive or negative, depending on how certain groups are viewed within a society. These biases arise when the reality of the world does not align with an envisioned ideal, leading to skewed perceptions and treatment of different groups. Historical bias is a prime example, where longstanding inequalities and stereotypes become ingrained in AI systems through the data they are trained on.

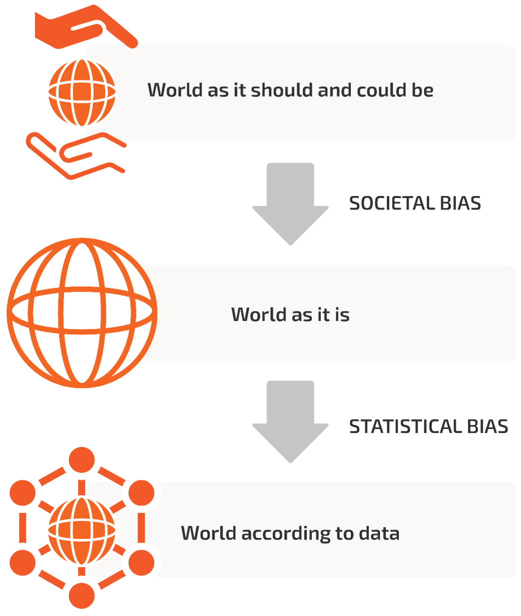


Figure 1: Societal and statistical biases

Statistical Bias is defined as a systematic difference between an estimated parameter in the data and its true value in the real world (European Union Agency for Fundamental Rights, 2022). This type of bias occurs when the data fails to accurately capture the intended variables or phenomena, leading to flawed AI outcomes. Examples of statistical bias include representation bias, where certain groups are underrepresented in the data, and measurement bias, where the variables used do not accurately reflect the concepts they are intended to measure.

In addition to these overarching categories, **cognitive biases** also play a significant role in AI development. Cognitive biases are systematic errors in thinking that can affect judgment and decision-making (Haselton et al., 2015). One common example is **confirmation bias**, where individuals tend to seek out or give more weight to data that confirms their pre-existing ideas or hypotheses, while disregarding information that contradicts these beliefs (Nickerson, 1998). Confirmation bias can be present at any stage of the AI lifecycle, making it essential to consciously strive for impartiality and objectivity.

To mitigate these biases, it is crucial to be aware of their potential influence and to implement strategies that promote fairness and accuracy throughout the AI development process. By understanding and addressing both societal and statistical biases, as well as remaining vigilant against cognitive biases, we can work towards developing AI systems that are more equitable and just.

FAIRNESS VS DISCRIMINATION

While discrimination and fairness are inherently linked to the concept of bias, it is crucial to recognize that they are distinct from both each other and bias itself. Bias refers to a systematic difference in the treatment of certain individuals or groups, without necessarily implying whether this difference is ‘right’ or ‘wrong.’ In contrast, discrimination and fairness introduce a value judgment regarding the outcomes of biased treatment.

A biased AI system can produce results that may be deemed ‘discriminatory’ or ‘unfair,’ depending on the context and the values applied. Understanding the distinctions between bias, discrimination, and fairness is essential for clear communication and effective action. Being aware of these differences helps ensure that when these concepts are discussed, everyone involved has a shared understanding of what is being referred to and the implications it carries.

DISCRIMINATION

Discrimination involves treating people differently, disadvantaging, or excluding them based on certain (personal) characteristics. In the Netherlands, discrimination is explicitly prohibited under Article 1 of the Constitution, which mandates equal treatment for all individuals. Dutch law further identifies specific protected attributes and clarifies how discrimination is categorized. The following attributes are identified as protected under Dutch law:

- RACE AND COLOR
- SEXUAL ORIENTATION
- GENDER
- POLITICAL OPINION
- NATIONALITY
- RELIGION AND BELIEF
- DISABILITY OR CHRONIC ILLNESS
- AGE
- MARITAL AND CIVIL STATUS

Internationally, other attributes are also recognized as protected grounds, such as ethnic or social origin, pregnancy, genetic features, language, membership of a national minority, property, and birth. Article 19 of the Treaty on the Functioning of the European Union establishes a specific set of protected grounds, including sex, racial or ethnic origin, religion or belief, disability, age, and sexual orientation. These grounds have formed the foundation for the adoption of EU Directives focused on ensuring equal treatment across member states.

At the national level, certain European countries, such as the Netherlands, have expanded their lists of protected attributes to cover more areas than those specified in the Treaty. However, despite these efforts, the limitation of the current lists of protected grounds, which excludes certain characteristics, results in a considerable gap, and many individuals who face discrimination still remain outside the scope of existing anti-discrimination laws in Europe.

The report from the European Network Equality Bodies (Equinet, 2021) highlights the need to expand the list of protected grounds in anti-discrimination law to provide protection for vulnerable social groups, prevent gaps in legal coverage, and reduce the burden on courts to interpret less obvious cases. It emphasizes the importance of recognizing socio-economic disadvantage, health status, and gender identity-related grounds to enhance legal protections and it also suggests considering additional grounds, such as genetic heritage and physical appearance, to ensure more inclusive coverage.

Besides the protected attributes, other statutory provisions, such as Article 1.1 of the Dutch 'Equal Treatment Law' (Wet AwGB, 2020), differentiate between direct and indirect discrimination:

- **Direct discrimination** occurs when a person is treated differently than someone else in a comparable situation, based on the protected attributes listed above.

- **Indirect discrimination** arises when a seemingly neutral provision, criterion, or practice disproportionately impacts individuals with a certain protected attribute.

While it is generally forbidden to differentiate based on attributes such as gender, there are exceptions when such differentiation is relevant to the situation. For example, organizing an all-female soccer tournament is permissible as it is specifically intended to exclude men. Additionally, positive discrimination—such as favoring candidates with diverse backgrounds when they are equally qualified—is allowed to promote diversity within the workforce.

Discrimination law presents significant challenges in practice, especially when it comes to proving discriminatory outcomes produced by AI systems. Individuals often lack access to the AI systems in question and may not have the technical knowledge needed to demonstrate that discrimination has occurred. This lack of transparency makes it difficult for both individuals and organizations to prove instances of discrimination.

Given these challenges, it is crucial for companies to proactively review and monitor their AI systems to prevent discriminatory outcomes. Implementing thorough checks and placing a strong emphasis on fairness can help mitigate the risk of discrimination and ensure that AI systems comply with legal and ethical standards.

FAIRNESS

While the definition of discrimination may appear straightforward in theory, the concept of fairness is much more complex, with multiple definitions and interpretations. Fairness generally revolves around the idea of being just, but what constitutes justice can vary significantly depending on cultural perspectives and the specific context. These varying interpretations make it challenging to establish a universally accepted definition of fairness (Mehrabi et al., 2021).

In the context of AI, unfairness can be understood as the "unjustified differential treatment that preferentially benefits certain groups over others" ("ISO/IEC 22989:2022," 2022). Fairness, therefore, is the absence of such unjustified differential treatment or prejudice toward any individual or group. The term 'algorithmic fairness' is often used interchangeably with fairness in AI. From a technical standpoint, fairness in an algorithm is achieved when it operates without being altered or manipulated for purposes unrelated to the users' interests (Varona and Suarez, 2022). This ensures that the algorithm genuinely serves the users' needs without being skewed by external influences that could introduce bias (Mehrabi et al., 2021).

Research on fairness in AI frequently focuses on developing metrics and tools to audit systems for biases. There are numerous fairness metrics, each with a different approach. Some metrics emphasize individual fairness, where the system should deliver similar predictions for similar individuals, while others focus on group and subgroup fairness, ensuring that different groups are treated equally (Mehrabi et al., 2021). One might argue that optimizing across all these metrics would yield the fairest system. However, this is not feasible because of the inherent mathematical tension between different fairness definitions. Optimizing one metric often comes at the expense of another (Ruf & Detyniecki, 2021). Therefore, it is crucial to determine which fairness metric best aligns with the goals of your AI system during its development.

Fairness and discrimination are deeply interconnected, with a strong interdependency between the two. Discrimination can be seen as a source of unfairness, particularly when it arises from differentiating based on sensitive attributes. In essence, fairness in AI aims to prevent discrimination by ensuring that systems produce unbiased outcomes that treat all individuals and groups equitably.

The Life Cycle of an AI System

Bias does not only manifest in the output of an AI system; it can emerge at various stages throughout the system's lifecycle. By adhering to a lifecycle model framework, we can systematically identify where and how some of the most significant biases may arise. This approach ensures that stakeholders involved in each phase of the AI lifecycle are aware of these potential biases and are equipped to take proactive measures to mitigate them effectively.

The lifecycle of an AI system outlines the interative journey from its inception to its eventual retirement, detailing the various stages it undergoes along the way. While there are different AI lifecycle frameworks available, this document uses as reference the model specified in ISO/IEC 22989 Information technology — Artificial intelligence — Artificial intelligence concepts and terminology. The reason for this choice is that this ISO standard provides standardized concepts widely recognized and used by a broad range of stakeholders and additionally, standardization also plays a crucial role in demonstrating conformity with regulations like the EU AI Act.

Making use of a lifecycle model framework offers stakeholders a structured approach to building AI systems with greater effectiveness and efficiency. This frameworks help manage the complexities inherent in AI development, deployment, and maintenance, ultimately leading to the creation of more robust and adaptive AI solutions that can meet diverse needs and challenges.

The ISO AI system lifecycle consists of eight stages, with the 'Design' and 'Development' phases further divided into subphases, as shown in Figure 2. This division was made due to the extensive range of activities occurring within these phases. By breaking it down into subphases, it becomes easier to identify and address the specific biases relevant to each part of the process.

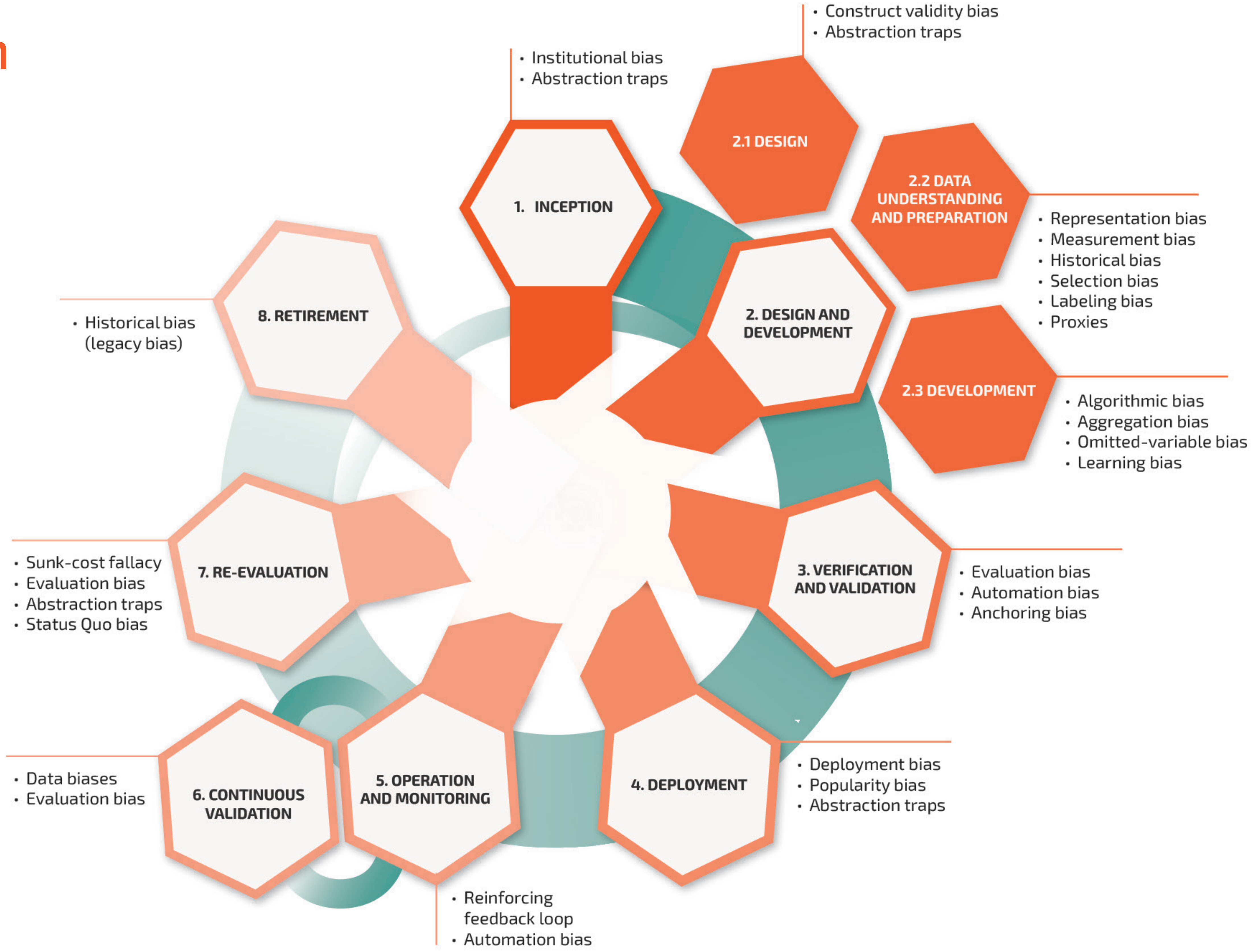


Figure 2: Bias per AI lifecycle Phase

BIAS AND MITIGATIONS PER AI LIFECYCLE PHASE

The development of an AI system is typically an iterative process. This means that certain stages may need to be revisited or repeated as the system evolves and new challenges arise. For instance, phases such as 'Design' and 'Development' and 'Deployment' might be revisited to address bugs, incorporate updates, or enhance the system's functionality. When re-evaluation leads to significant changes, it may be necessary to loop back to earlier stages like 'Inception' or 'Design' and 'Development' to ensure that the system is aligned with its original goals and requirements. Moreover, there is a continuous loop between the 'Operation and Monitoring' and 'Continuous Validation' phases, especially in AI systems that employ continuous learning where regular assessment and validation are crucial to maintain accuracy and performance over time.

1. INCEPTION

The inception phase begins when stakeholders commit to transforming an idea into a functional AI system. This stage involves making key decisions and developing plans that will guide the project into the subsequent design and development phase. However, this phase may need to be revisited if new information arises indicating that the system is not feasible or financially viable.

2.1 DESIGN

In the design phase, the conceptual framework of the AI system is established. This includes determining the system's architecture, selecting the appropriate algorithms, and identifying the resources needed for successful implementation.

2.2 DATA UNDERSTANDING AND PREPARATION

In this phase, relevant datasets are identified and analyzed to assess their quality, structure, and potential relevance. Following this analysis, the data is collected, cleaned, and prepared for the subsequent model development phase.

2.3 DEVELOPMENT

The development phase is the stage where the core AI model is built and iteratively refined. This phase involves selecting appropriate algorithms and applying them to the prepared datasets to create a functioning model. Typically, the dataset is split into training and testing sets. Throughout this phase, developers may adjust model parameters, optimize algorithms, and validate the model's outputs to ensure it meets the desired accuracy and performance standards.

3. VERIFICATION AND VALIDATION

After the model is developed, it undergoes validation to ensure that it performs effectively on new and unseen data, in line with the established requirements and objectives. This validation process includes evaluating the model's accuracy, performance, and overall reliability to confirm that it meets the intended goals and is ready for deployment.

4. DEPLOYMENT

During the deployment phase, the AI system is implemented and integrated into the business process. At this stage, the system becomes fully operational and begins processing real data in a live environment, functioning as intended to support business objectives.

5. OPERATION AND MONITORING

The AI model is now running and available for use, with continuous monitoring to ensure it maintains expected performance levels. During this phase, the system is supported with necessary repairs, updates, and assistance to both the AI system and its users.

6. CONTINUOUS VALIDATION

This stage refers to the ongoing process of monitoring and evaluating an AI system's performance while it is operational, particularly in systems that employ continuous learning. In this phase, the AI model undergoes incremental training using new data as it becomes available, allowing the system to improve over time. Continuous validation ensures that the AI system maintains its accuracy, reliability, and alignment with its intended goals.

7. RE-EVALUATION

The re-evaluation phase may take place at predefined intervals, in response to significant changes in the environment or in performance. The primary goal of this phase is to ensure that the AI system continues to meet its objectives and address any newly identified risks. This phase is not a formal stage in all AI lifecycle models but is important in iterative development processes or when continuous learning is involved.

8. RETIREMENT

When the AI model is no longer needed or is replaced by an improved version, it is archived. This process includes retaining all relevant documentation and associated artifacts to ensure important information is preserved for future reference, auditing, or compliance purposes.

LIFE CYCLE OF AN AI SYSTEM

PHASE 1

INCEPTION

TYPES OF BIAS IN THIS PHASE

- Institutional bias
- Abstraction traps

The inception phase is the first stage in the AI development lifecycle and begins when stakeholders decide to transform an idea into a real system. Stakeholders may be eager to pursue an idea where they believe an AI model is the ideal solution. However, what many may not realize is that biases can emerge right from the outset, potentially influencing the entire development process.

Early decisions made during the Inception phase can reflect systemic biases present within organizational settings, individuals, and groups, leading to decisions that are shaped by a narrow perspective. These initial decisions can significantly impact subsequent stages, potentially resulting in biased outcomes. Two primary biases that can emerge during the inception phase are Institutional Biases and Abstraction Traps.

INSTITUTIONAL BIAS

Institutional bias refers to systemic tendencies within entire institutions that result in favoring or disadvantaging specific social groups (Schwartz et al., 2022). Unlike biases that occur at the individual level, institutional bias operates at the organizational level, where established practices or norms perpetuate unequal treatment. This can manifest in various forms, such as discriminatory hiring practices, unequal access to resources, or disparities in opportunities.

Institutional biases, such as institutional racism and institutional sexism, are deeply embedded within organizational structures and policies, often reinforcing systemic inequalities. The decisions made at the start of developing an AI system are crucial, as they shape its eventual outcomes by determining who and what is considered or excluded. This early influence can lead to biased results that reflect and perpetuate these institutional biases.

Recommendations for addressing institutional bias

- **Identify and involve diverse stakeholders**: Clearly outline the stakeholders who should be involved in each phase of the AI lifecycle. Different stakeholders bring unique expertise and experiences, making it crucial to involve them throughout the process (Muhammad, 2022). Ensuring that stakeholders represent diverse perspectives and backgrounds can help mitigate the risk of overlooking blind spots that may arise from homogeneous groups.
- **Define affected demographic groups**: It is essential to identify and define the demographic groups that the AI system is likely to impact (Muhammad, 2022). Understanding who the system will affect helps ensure that the needs and concerns of these groups are considered, reducing the risk of institutional bias and promoting more equitable outcomes.

ABSTRACTION TRAPS

When translating a real-world problem into an AI system, the process inherently involves simplifying reality into a model. This simplification requires the removal of certain details, which can lead to unintended consequences if critical contexts are abstracted away.

Abstraction traps refer to the pitfalls that arise when the social context surrounding an AI model and its inputs and outputs are oversimplified or ignored. This failure to adequately consider the interplay between technology and its social environment can result in significant oversights (Selbst et al., 2019). Key abstraction traps include:

- **The Formalism Trap**: This occurs when the formulation of a problem for an AI model neglects to sufficiently consider the context in which the model will be applied. By focusing too narrowly on the technical aspects, important social, ethical, or practical considerations may be overlooked.

- **The Ripple Effect Trap**: This trap involves a failure to recognize how introducing an AI system into an existing social system can alter the behaviors of other actors within that system. These changes can, in turn, reshape the context in which the AI operates, potentially leading to unintended and unforeseen consequences.
- **The Solutionism Trap**: This occurs when there is a failure to acknowledge that the best solution to a given problem may not involve the use of an AI system. Over-reliance on AI can lead to neglect simpler, more effective, or more ethical solutions that do not require advanced technology (Selbst et al., 2019; Weerts, 2021).

Recommendations for addressing abstraction traps

The Formalism Trap

- **Align problem formulation with social context:** Stakeholders should ensure that the problem formulation accurately reflects the understanding of relevant social constructs within the intended deployment context (Weerts, 2021). This alignment is crucial to avoid oversimplifying complex social dynamics and to ensure the AI system is appropriately tailored to the environment in which it will be applied.

The Ripple Effect Trap

- **Consider the broader impact on all actors:** It is important to pay close attention to how the introduction of the AI system might affect the behavior, perception, and expertise of all actors involved, not just those who directly interact with the model. This includes considering individuals whose work or roles may be indirectly impacted by the system.
- **Analyze power dynamics**: Investigate the existing power dynamics among the actors within the system to anticipate any shifts that might occur due to the introduction of new technology (Muhammad, 2022). Understanding these dynamics can help prevent unintended consequences and ensure a more equitable deployment of the AI system.

The Solutionism Trap

- **Critically evaluate the need for AI:** While AI is often seen as a solution to many problems, it is essential to critically evaluate whether AI is truly the most appropriate tool for addressing the specific issue at hand. Before deciding to develop an AI system, stakeholders should consider whether the problem can be effectively addressed through alternative approaches that do not rely on AI. This careful consideration helps avoid unnecessary complexity and ensures that the chosen solution is both efficient and ethically sound.

Example: The Solutionism Trap in Education

In education, the solutionism trap can occur when schools invest heavily in AI-driven personalized learning platforms without adequately considering the importance of teacher-student interactions and individualized instruction. While AI can effectively tailor learning materials to students' preferences and abilities, it may overlook the social and emotional aspects of learning that are crucial for student engagement and academic success. This oversight can lead to a reliance on technology at the expense of the holistic development that personal interaction fosters.

PHASE 2.1 DESIGN

TYPES OF BIAS IN THIS PHASE

- Construct validity bias
- Abstraction traps

After the 'Inception' phase, stakeholders move on to the design stage, where they outline the approach for developing and testing their AI system. During this phase, critical decisions are made regarding the system's design architecture, including whether to develop custom hardware and software, purchase existing solutions, or utilize open-source resources. These choices significantly impact the potential biases that may arise during the system's development. Of particular concern at this stage are construct validity bias and additional abstraction traps.

It is also important during the design phase to begin considering which fairness definitions and metrics will be applied, and why these choices are appropriate for the specific AI system being developed. Making informed decisions about fairness at this stage lays the foundation for ensuring that the system operates equitably and aligns with its intended objectives.

CONSTRUCT VALIDITY BIAS

Construct validity bias is a type of statistical bias that occurs when a feature or target variable fails to accurately measure the construct it is intended to represent (Weerts, 2021). This bias is particularly common when dealing with complex or abstract concepts that are difficult to quantify. For example, socioeconomic status is a multifaceted construct that might be partially measured by income, but income alone does not account for other critical factors such as wealth and education (Jacobs & Wallach, 2021).

It's important to note that construct validity bias is not confined to the design phase; it can arise during other phases of the AI lifecycle as well. Recognizing and addressing this bias early on is essential for developing AI systems that more accurately reflect the constructs they are designed to measure.

Recommendations for addressing construct validity bias

- **Collect multiple measures for complex constructs**: To mitigate construct validity bias, it is advisable to collect multiple measures for complex constructs (Weerts, 2021). This approach helps enhance the comprehensiveness and accuracy of the measurement, ensuring that the construct is more fully represented in the AI system.
- **Document and report considerations for target variables and features**: It is crucial to report the considerations made for the target variable and any (sensitive) features used in the AI system (Muhammad, 2022). Specifically, describe how these variables are measured and the rationale behind their selection. This transparency helps to clarify the assumptions made and the potential limitations in how these variables are represented.
- **Acknowledge variability in interpretation of features**: Recognize that certain features can have different meanings for different individuals. For example, socioeconomic status (SES) is a complex construct that may include various indicators such as income, wealth, education level, and occupation. However, individuals may interpret and define SES differently based on their cultural background, upbringing, or personal experiences. Understanding and accounting for this variability is essential to avoid oversimplification and to ensure a more accurate representation of the construct in the AI system.

ABSTRACTION TRAPS

As previously highlighted, abstraction traps occur when the social context connected to technology is overlooked because important details surrounding an AI model and its inputs and outputs are abstracted away (Selbst et al., 2019). During the design phase, there is a risk of encountering abstraction traps (Selbst et al., 2019; Weerts, 2021) such as:

- **The Framing Trap**: This trap arises when there is a failure to adequately model the broader context or relevant aspects of the larger system in which an AI system operates. AI systems are often integrated into decision-making processes that involve other systems or human decision-makers. If the relevant criteria of the larger system are not fully accounted for, the AI model may produce inaccurate or incomplete outcomes. The framing trap is closely related to construct validity, as both involve the risk of oversimplifying or misrepresenting complex realities.
- **The Portability Trap**: This trap occurs when there is a failure to recognize that repurposing an AI system designed for one context may lead to inaccuracies or unintended harm when applied to a different context. This issue can arise due to shifts in geographical location, time, domain, or other contextual factors. Applying a model outside its original design parameters without proper adjustments can lead to significant errors and unintended consequences.

Recommendations for addressing construct abstraction traps

The Framing Trap

- **Assess problem framing and solution evaluation**: Ensure that when framing the problem and evaluating the solution, all relevant components and actors within the sociotechnical system are considered (Weerts, 2021). This holistic approach helps to account for the broader context in which the AI system will operate, reducing the risk of overlooking critical factors.

The Portability Trap

- **Consider contextual differences**: Evaluate how factors such as geographical location, cultural norms, and temporal changes might influence the performance and suitability of the AI system. When reusing a model in a different context, clearly identify the differences between the original and new contexts and assess how these variations may affect the model's outcomes (Weerts, 2021).

Example: The Portability Trap in Self-Driving Cars

In the development of self-driving cars, an AI system might be trained to navigate urban environments using data collected from a specific country. However, when the system is deployed in a different country with unique traffic patterns, road infrastructure, and driving behaviors, it encounters the portability trap. The AI system may struggle to adapt to these new conditions, leading to unsafe driving behaviors or an increased risk of accidents. Addressing the portability trap requires extensive testing and fine-tuning to ensure the model performs safely and effectively across diverse geographical locations.

PHASE **2.2**

DATA UNDERSTANDING AND PREPARATION

TYPES OF BIAS IN THIS PHASE

- Representation bias
- Measurement bias
- Historical bias
- Selection bias
- Labeling bias
- Proxies

In the 'Data Understanding and Preparation' phase of developing an AI system, data are either collected or datasets are selected for use. Thorough exploration and preprocessing of the data are essential steps, as the quality and integrity of the data directly influence the performance and reliability of the resulting model. The saying "garbage in, garbage out" underscores the critical importance of high-quality data inputs.
However, this phase is also where many biases are likely to emerge, stemming from both statistical and societal factors. This makes the 'Data Understanding and Preparation' phase particularly crucial. Biases at this stage can arise from various sources, including lack of representation or selection bias, historical biases, measurement biases, and more. Identifying and addressing these biases early on is essential to developing a fair and accurate AI system.

REPRESENTATION BIAS

Representation bias occurs when the data does not accurately reflect the diversity or complexity of the population it is intended to model. This can lead to the AI system failing to generalize well to the real-world use population, resulting in increased errors, particularly for minority groups (Mehrabi et al., 2021).
Representation bias can arise in several ways (Suresh & Guttag, 2021):

- If the defined target population does not accurately represent the use population, the resulting model may be biased.
- If the target population includes underrepresented groups, the model may not perform well for these groups.
- If the sampling method used to collect data from the target population is limited or uneven, it can lead to a biased dataset that does not fully capture the diversity of the population, leading to skewed model performance.

Recommendations for addressing representation bias

- **Ensure balanced representation**: Verify that your dataset includes a balanced representation of all subgroups present in the model, with a particular focus on ensuring there are sufficient instances of minority groups (Van Giffen et al., 2022). If the dataset is imbalanced, consider collecting additional data to address this issue. Applying data visualization techniques can also be beneficial for gaining insights into the distribution and representation of different groups within your dataset.
- **Apply sampling techniques**: To achieve a balanced dataset, you can use sampling techniques such as oversampling, undersampling, and stratified sampling. These methods should be applied exclusively to the training set to avoid introducing bias into the model evaluation process. The validation and test sets should remain unaltered to ensure they accurately reflect the real-world population and provide an unbiased assessment of the model's performance (Muhammad, 2022).

MEASUREMENT BIAS

Measurement bias occurs when there is a systematic or non-random error in data collection that causes errors to be greater for some groups than for others (Mehrabi et al., 2021). This type of bias can significantly impact the accuracy and fairness of an AI system, particularly when certain features are involved.

Measurement bias becomes more problematic under the following conditions (Suresh & Guttag, 2021):

- When the accuracy of the measurement varies among different groups, it can lead to biased outcomes, with some groups experiencing higher error rates than others.
- If the measurement method used differs across groups, this can introduce inconsistencies that result in biased data.
- When a feature used in the model is an oversimplification of a more complex construct, it may fail to capture the nuances of that construct accurately, leading to biased conclusions.

Recommendations for addressing measurement bias

- **Re-evaluate the measurement process**: Carefully re-examine the measurement process by considering the context in which the data is collected and critically assessing how the data is measured or annotated. This involves scrutinizing the methods used and identifying potential biases inherent in the process (Muhammad, 2022).
- **Collaborate with domain experts**: Work closely with domain experts to exchange knowledge and insights. Their expertise can provide a deeper understanding of the underlying causes of measurement bias and help identify more accurate proxies for variables of interest, ensuring that the data collected is both relevant and fair (Van Giffen et al., 2022).

HISTORICAL BIAS

Historical bias occurs even when data are perfectly measured and sampled, as it arises from the inherent biases embedded in the world as it exists or existed. This type of bias reflects the existing social and cultural inequalities that have been historically present, meaning that even if an AI system accurately mirrors the real world, it may still perpetuate or amplify harm to certain populations (Suresh & Guttag, 2021).

Historical biases can also contribute to other types of biases, such as construct validity bias or labeling biases, particularly when labels are based on human judgment (Muhammad, 2022). These biases are often deeply rooted in past practices, norms, or structures, making them challenging to identify and mitigate without a conscious effort to understand the historical context in which the data was generated.

Recommendations for addressing historical bias

- **Improve the representation of minority groups**: Since historical bias is often caused by inadequate representation of minority groups in the dataset, enhancing the representation of these groups can help mitigate this bias. This could involve using over- and undersampling techniques, as well as revisiting strategies to address representation bias (Muhammad, 2021).
- **Collaborate with domain experts**: Work closely with domain experts to identify and analyze any unjust patterns embedded in the dataset. By exchanging insights with these experts, you can ensure that relevant and measurable features are included, which helps to address and reduce the impact of historical bias (Van Giffen et al., 2022).

Example: Historical Bias in Healthcare AI

Imagine an AI-powered healthcare diagnostic tool trained on historical patient data. In this dataset, patients from minority racial groups are underrepresented due to longstanding disparities in healthcare access and systemic barriers. This underrepresentation is rooted in historical biases, such as discriminatory practices or socioeconomic inequalities, which have restricted healthcare resources for these communities. As a result, the tool lacks sufficient data on conditions that are prevalent in these groups, leading to less accurate diagnoses for minority patients. Even though the dataset does not include explicit racial features, the AI system's outcomes exhibit racial bias due to the historical biases embedded in the data.

SELECTION BIAS

Selection bias is a statistical bias that occurs when data collection or selection procedures result in a non-random sample of the population (Mehrabi et al., 2021). This bias is closely related to representation bias, and there are several specific types of biases that fall under the category of selection bias (Muhammad, 2022):

- **Sampling bias**: This occurs when data is not randomly collected from the target group, leading to a sample that may not be representative of the entire population.
- **Self-selection bias**: This arises when certain groups of people are more likely to opt out of the data collection process, leading to an underrepresentation of these groups in the dataset.
- **Coverage bias**: Coverage bias happens when the population that we aim to make predictions about is not accurately represented in the dataset. This can lead to skewed predictions and unreliable model outcomes.

Recommendations for addressing selection bias

Selection bias often arises from the underrepresentation of certain demographic groups in the data (Muhammad, 2022). To mitigate selection bias, it is crucial to gather diverse and representative datasets. This involves sourcing data from a wide range of demographics to ensure inclusivity and fairness in representation.

To address selection bias effectively, it is helpful to revisit the recommendations provided for mitigating representation bias. Additionally, consider asking critical questions such as:

- **How were the data samples selected, and what criteria were used?**
- **Have alternative sampling methods been considered?**

Example: Selection Bias in Surveys

For instance, consider a survey conducted to determine why individuals did not vote in an election. If the survey was carried out at the exit of a shopping mall, it would exclude people who do not frequently visit that specific area or do not shop at malls. This could lead to selection bias in the data sample, as the responses collected may not accurately represent the broader population, particularly those who were not at the mall.

LABELING BIAS

Labeling bias can occur when the data includes labels assigned by annotators, and these annotators may have different interpretations of the same label (Jiang & Nachum, 2019). This inconsistency can lead to biases in the dataset, which may negatively impact the accuracy and fairness of the AI model.

Recommendations for addressing labeling bias

- **Define clear labeling requirements**: It is crucial to establish clear labeling requirements from the start. Outline the specific classes to be labeled and clearly define the responsibilities of the annotators for each dataset. Collaborate with domain experts to gain insights that can help reduce ambiguity in these decisions, ensuring that the labels are accurate and meaningful (Van Giffen et al., 2022).
- **Determine labeling methods and performance metrics**: Decide which tasks require manual labeling and which can be handled through automated annotation. Additionally, establish how you will measure the performance of these labeling processes. To maintain consistency and objectivity, consider asking questions such as:
  - **Are there subjective interpretations in the labeling process?**
  - **Have measures been taken to ensure consistency and objectivity in labeling?**

PROXIES

A proxy variable is a substitute or indirect measure used to represent a concept or construct that is difficult to directly observe or quantify (Muhammad, 2022). While proxies can be useful, they pose a risk when there is an underlying correlation with a sensitive attribute, potentially leading to indirect discrimination (Borgesius, 2018). As Barocas and Selbst (2016) explain, “Criteria that are genuinely relevant in making rational and well-informed decisions also happen to serve as reliable proxies for class membership.” This means that even when datasets do not explicitly contain sensitive features, proxies can still result in differential treatment of certain groups. Therefore, it is crucial to carefully evaluate and manage the use of proxies to prevent unintended bias in AI systems.

Recommendations for addressing proxies

- **Select the appropriate proxies**: When direct measurement of variables of interest is not possible, it is often necessary to choose proxies. However, it is essential to critically examine the underlying correlations between these proxies and the true variables of interest. (Van Giffen et al., 2022). Understanding whether a proxy merely correlates with a sensitive attribute or actually reflects a causal relationship helps in selecting appropriate proxies that do not inadvertently lead to bias.
- **Consider sensitive features**: Collecting sensitive features can be crucial in developing fair algorithms (Žliobaitė & Custers, 2016). The absence of sensitive features makes it challenging to assess whether an AI model might become discriminatory or unfair. By carefully evaluating the relationship between proxies and sensitive attributes, you can better ensure that the AI system operates fairly and avoids indirect discrimination.

PHASE 2.3 DEVELOPMENT

TYPES OF BIAS IN THIS PHASE

- Algorithmic bias
- Aggregation bias
- Omitted-variable bias
- Learning bias

During the 'Development' phase, selected models are constructed, trained on the prepared dataset, and optimized to address the original problem. This phase is crucial, as it is particularly susceptible to various biases that can affect the effectiveness and fairness of the AI system. These biases include algorithmic bias, aggregation bias, omitted-variable bias, and learning bias. Addressing these biases during the 'Development' phase is essential to ensure that the resulting AI system operates fairly and meets its intended objectives.

ALGORITHMIC BIAS

Algorithmic bias refers to a type of bias that is introduced by the algorithm itself, rather than being present in the dataset. This bias can arise from various design decisions made during the algorithm's development, such as the selection of optimization functions, the application of regularization in regression models, and the use of statistically biased estimators within algorithms (Mehrabi et al., 2021). These decisions, if not carefully considered, can lead to unintended biases that affect the fairness and accuracy of the AI system.

Recommendations for addressing algorithmic bias

- **Employ de-biasing techniques**: Use de-biasing techniques to enhance the accuracy and fairness of predictions. These methods involve adjusting the model or dataset to mitigate biases, thereby improving the reliability and equity of the AI system's outcomes (Nazer et al., 2023).
- **Ensure transparency, interpretability, and reproducibility**: Promote transparency, interpretability, and reproducibility in the model's methodology. Providing clear explanations of the algorithm's workings and decision-making processes helps stakeholders understand and trust the model's outcomes. This transparency is crucial for identifying and mitigating biases within the algorithm (Nazer et al., 2023; Van Giffen et al., 2022).
- **Incorporate fairness constraints**: Consider introducing regularization terms or constraints that account for differences in how the learning algorithm classifies protected and non-protected groups. By integrating fairness constraints into the model's training process, the algorithm can learn to make predictions that are more equitable across different demographic groups (Van Giffen et al., 2022).

AGGREGATION BIAS

Aggregation bias occurs when a one-size-fits-all model is applied to groups that actually have distinct data distributions. This bias arises from the incorrect assumption that the data distribution is uniform and homogeneous across all groups, leading to misleading conclusions. In reality, different underlying groups may require separate consideration due to their unique characteristics and data patterns (Suresh & Guttag, 2021; Mehrabi et al., 2021). This can result in the AI system producing suboptimal or biased outcomes for certain groups, as it fails to account for these differences.

Recommendations for addressing aggregation bias

- **Incorporate group differences into the objective function**: Modify the objective function to account for differences between groups within the data (Suresh & Guttag, 2021). This adjustment can help the model better recognize and learn from the distinct data distributions, potentially improving performance across different demographic groups.
- **Address underfitting**: Aggregation bias is often related to underfitting, where certain model classes fail to adequately capture the varying data distributions among different groups. It's important to recognize this risk (Muhammad, 2022). To mitigate underfitting and reduce aggregation bias, consider increasing the sample size for underrepresented groups, ensuring that the model has sufficient data to learn effectively from these groups.

OMITTED-VARIABLE BIAS

Omitted-variable bias occurs when a relevant feature that influences both the independent and dependent variables is excluded from the statistical model. This omission can lead to distorted estimates of the relationships between the included features, as the model fails to account for an important factor that could affect the outcomes (Mehrabi et al., 2021). This bias can significantly impact the accuracy and validity of the model's predictions, leading to misleading conclusions.

Recommendations for addressing omitted-variable bias

- **Apply feature importance methods**: Use feature importance methods to evaluate the relationship between each feature and the target variable (Muhammad, 2022). This approach can help identify and ensure that relevant features are included in the model, reducing the risk of omitting variables that could significantly impact the model's accuracy and the validity of its predictions.

LEARNING BIAS

Learning bias occurs when a model prioritizes one objective, such as accuracy, at the expense of another, such as a fairness-related metric (Suresh & Guttag, 2021; Muhammad, 2022). This bias emerges during the training process, where the model may inadvertently favor certain outcomes that optimize performance for one metric, while compromising on other important aspects like fairness, leading to imbalanced or unfair results

Recommendations for addressing learning bias

- **Critically select optimization metrics**: Thoughtfully consider which metrics you choose to optimize. Include subgroup metrics alongside overall performance metrics to monitor variations in the model's performance across different groups. This approach helps ensure that the model does not disproportionately favor certain objectives, such as accuracy, at the expense of fairness.
- **Address representation bias**: Learning bias can exacerbate accuracy differences for underrepresented groups. To mitigate this, it is crucial to address representation bias by ensuring a more representative and balanced dataset. By doing so, the model is less likely to disproportionately learn from the majority group data, leading to more equitable outcomes across all demographic groups (Muhammad, 2022).

Example: Omitted-variable Bias in Subscription Service Prediction

Consider a scenario where a model is developed to predict the annual percentage rate of customers who might cancel their subscription to a service. The model achieves high accuracy in its predictions, but it fails to anticipate a sudden surge in cancellations. This surge is later attributed to the appearance of a new, lower-priced competitor in the market— a factor that the model did not account for. By omitting this crucial variable, the model's predictions were significantly distorted, leading to an incomplete understanding of customer behavior (Mehrabi et al., 2021).

PHASE 3

VERIFICATION AND VALIDATION

TYPES OF BIAS IN THIS PHASE

- Evaluation bias
- Automation bias
- Anchoring bias

During the 'Verification and Validation' phase of AI system development, rigorous checks are conducted to ensure that the system functions according to the specified requirements and achieves its intended objectives. Verification focuses on testing the software and hardware components for functionality, identifying bugs, and assessing integration, while performance tests evaluate the system's response time and other critical characteristics.

A key aspect of this phase is verifying that the AI system's capabilities operate as intended, which requires the acquisition, preparation, and use of representative test data that is separate from the development data. This helps ensure that the system is tested under conditions that closely resemble real-world scenarios.

Additionally, stakeholders assess the system's functional completeness and overall quality to determine whether it is ready for deployment. However, this phase is also susceptible to various biases, including evaluation bias, automation bias, and anchoring bias, which can influence the testing and validation processes, potentially leading to flawed assessments of the system's performance and readiness.

EVALUATION BIAS

Evaluation bias arises when the metrics and procedures used to evaluate a model's performance are not appropriately aligned with the model, dataset, or the population on which the model will be deployed. This misalignment can result in misleading assessments, as the evaluation may not accurately reflect how the model will perform in real-world scenarios (Suresh & Guttag, 2021). This bias can undermine the effectiveness and fairness of the AI system if not properly addressed during the validation process.

Recommendations for addressing evaluation bias

- **Assess the suitability of evaluation metrics**: Critically evaluate the assumptions behind the chosen evaluation metrics to determine whether they are appropriate for your specific model and dataset (Muhammad, 2022). Ensuring that the metrics align with the model's intended use and the characteristics of the target population is essential for an accurate assessment.
- **Compare performance across groups**: Compare evaluation metrics across different subgroups to gain insights into how the model performs for various segments of the population (Van Giffen et al., 2022). This comparison can help identify disparities in performance that may affect certain groups, allowing for targeted improvements.
- **Mitigate overfitting**: Be aware of the risk of overfitting during the validation process, which can compromise the model's generalizability, particularly for underrepresented groups (Nazer et al., 2023). Implement strategies such as cross-validation, regularization, or the use of more robust evaluation techniques to mitigate overfitting and ensure the model performs well across diverse populations.
- **Monitor data distribution imbalances**: Regularly assess the data distribution for imbalances among subpopulations. If significant imbalances are identified, consider revisiting the 'Data Understanding and Preparation' phase to implement appropriate mitigation strategies (Muhammad, 2022). This ongoing monitoring helps maintain fairness and accuracy in the model's evaluations.

AUTOMATION BIAS

Automation bias is the tendency to favor suggestions or decisions made by AI systems, even when there are warning signals or conflicting information from other sources (Khera et al., 2023). This bias can lead to over-reliance on AI outputs, potentially ignoring critical inputs that could prevent errors. The consequences of automation bias can be severe, as illustrated by cases where human drivers have driven their cars into rivers after blindly following incorrect GPS routing instructions (Santiago, 2019). This example highlights the risks of unquestioningly trusting AI systems, underscoring the importance of maintaining human oversight and critical thinking when interacting with automated systems.

Recommendations for addressing automation bias

Awareness and training: Automation bias may not be present in all AI systems, but it can have significant implications when it does occur. For example, studies have shown that clinicians often favor automated decision-making systems, relying on AI tools even when they encounter contradictory or clinically nonsensical information (Khera et al., 2023). While this bias can benefit patients in cases where the AI model performs well, it can also pose risks in situations where the model is inaccurate, whether due to systematic bias or imperfect performance. In such scenarios, clinicians may defer to the AI model over their own judgment, potentially leading to harmful outcomes. To mitigate automation bias, it is essential that individuals who process or work with the results of AI systems receive proper training. They should be equipped to critically evaluate the outcomes generated by the system and maintain a skeptical, analytical approach to ensure that they do not blindly rely on AI outputs (Muhammad, 2022). This training can help prevent over-reliance on AI and promote better decision-making by integrating human expertise with AI insights.

ANCHORING BIAS

Anchoring bias is a cognitive bias that occurs when individuals rely too heavily on initial information or “anchors” when making decisions or judgments (Rastogi et al., 2022). It occurs when individuals fixate on a particular figure or course of action, which then influences how they interpret new information, leading to a distorted perception (Rastogi et al., 2022). This bias can make it difficult to adjust plans or decisions significantly, even when the situation warrants it.

Recommendations for addressing anchoring bias

For stakeholders involved in the ‘Verification and Validation’ phase, it is crucial to be aware of the potential for cognitive biases like anchoring bias to influence their work. By consciously taking the time and effort to avoid jumping to conclusions, stakeholders can ensure that the model validation process remains as objective and comprehensive as possible. This mindfulness helps in making well-informed decisions that reflect the true performance and suitability of the AI system.

Example: Anchoring Bias in the Verification and Validation Phase

In the ‘Verification and Validation’ phase, anchoring bias might lead stakeholders to form overly optimistic or pessimistic expectations based on early test results, which can influence subsequent assessments of the AI system’s performance. For instance, if initial performance tests yield exceptionally high or low scores, stakeholders may unconsciously adjust their acceptance criteria based on these early results. This can result in biased conclusions about the system’s readiness for deployment, as decisions are influenced by initial impressions rather than a comprehensive evaluation of all test outcomes.

PHASE 4
DEPLOYMENT

TYPES OF BIAS IN THIS PHASE

- Deployment bias
- Abstraction traps
- Popularity bias

In the 'Deployment' phase, the AI system transitions from development to being installed, released, or configured in its real-life environment. This stage introduces new challenges and biases that stakeholders must be mindful of. Deployment bias, abstraction traps, and popularity bias are among the biases that can emerge during this phase, potentially affecting the effectiveness and fairness of the AI system in real-world applications, with significant real-life consequences.

DEPLOYMENT BIAS

Deployment bias arises when there is a mismatch between the environment in which the AI system was developed and the environment in which it is ultimately deployed (Suresh & Guttag, 2021). These differences can include variations in data distributions, user behaviors, or system configurations, all of which can lead to unexpected performance issues or failures.

For example, if end users do not interact with the model as intended, the system's performance can become unpredictable, potentially undermining its effectiveness and reliability in the real-world setting. Recognizing and addressing deployment bias is crucial to ensure that the AI system functions as expected in its operational environment.

Recommendations for addressing deployment bias

- **Promote stakeholder discussions**: Stakeholders should continuously engage in discussions about the technical and social consequences of deploying an AI system. It is important to regularly reassess whether the deployed environment aligns with the system's intended function and to make adjustments as necessary (Van Giffen et al., 2022).
- **Enhance model interpretability and understandability**: Ensuring that deployed models have a high level of interpretability and understandability is crucial. A transparent model makes it easier to detect and address errors, which is particularly important for identifying issues related to both deployment bias and automation bias (Muhammad, 2022). By prioritizing these aspects, stakeholders can better monitor the system's performance and take corrective actions when needed, ensuring the AI system operates effectively in its real-world environment

ABSTRACTION TRAPS

As discussed in the 'Inception' and 'Design' phases, abstraction traps occur when the social context intertwined with technology is overlooked due to oversimplifying or abstracting away the context surrounding an AI model and its inputs and outputs (Selbst et al., 2019). During the 'Deployment' phase, abstraction traps—such as portability traps and framing traps—can exacerbate deployment bias, leading to unforeseen challenges in how the AI system operates in its new environment (Muhammad, 2022). These traps can result in the system being misaligned with the social and operational contexts in which it is deployed, potentially diminishing its effectiveness and fairness.

POPULARITY BIAS

Popularity bias occurs in recommendation systems when popular items are recommended more frequently than less popular ones, often at the expense of diversity and personalization (Naghiaei et al., 2022). This bias is specific to recommender systems and can lead to a feedback loop where already popular items continue to dominate recommendations, while niche or less-known items are overlooked, potentially reducing the overall effectiveness and fairness of the recommendation system.

Recommendations for addressing popularity bias

Developers should be mindful of popularity bias because it can significantly impact the fairness and effectiveness of their systems. Unlike other biases, popularity bias may not directly affect individuals, but it can influence the products and services they are exposed to, thereby impacting their daily lives. By being aware of this bias, developers can take steps to ensure that their AI systems provide a more balanced and diverse range of recommendations, rather than being overly influenced by trends or dominant perspectives. Mitigation measures can be applied at different stages of the AI lifecycle, each with its own techniques: (Klimashevskaia et al., 2024)

- **Pre-processing methods**: These are less commonly used and take place during the 'Data Understanding and Preparation' phase or the 'Development' phase. Methods include data sampling, item exclusion, or creating positive-negative sample pairs for learning.
- These techniques aim to reduce bias before the model is trained by balancing the data inputs.
- **In-processing methods**: These are more commonly applied during model training. Some approaches include:

  Regularization-based Approaches
  Constraint-based Approaches
  Re-Weighting Approaches
  Graph-based Similarity Adjustment
  Integration of Side Information
  Natural Language Processing-based Approaches
  Causal Inference-based Approaches
- **Post-processing methods**: These techniques adjust model outputs after training and include:

  Re-scaling (score adjustment)
  Re-ranking (reordering)
  Rank aggregation

Example: Deployment Bias in Sentiment Analysis

Consider a sentiment analysis model trained on social media data that performs well in a controlled lab environment. However, when deployed in a real-world setting, the model may struggle to accurately analyze sentiment. This could be because customers express their sentiments using different vocabulary, tones, or cultural references that the model was not exposed to during training. As a result, the model may fail to interpret these real-world expressions accurately, highlighting the presence of deployment bias.

PHASE 5 OPERATION AND MONITORING

TYPES OF BIAS IN THIS PHASE

- Reinforcing feedback loop
- Automation bias

During the 'Operation and Monitoring' phase, the AI system is actively deployed and accessible for use in real-world situations. This phase involves ongoing monitoring of the system to ensure it functions correctly and meets performance expectations. It also includes addressing any issues or errors that arise, updating software and hardware as needed, and providing support to users. In this phase, there is a continuous feedback loop with the 'Continuous Validation' phase, particularly if the AI system employs continuous learning. This loop ensures that the system adapts to new data and conditions while maintaining its effectiveness and reliability.

REINFORCING FEEDBACK LOOP & AUTOMATION BIAS

A Reinforcing Feedback Loop refers to feedback mechanisms that amplify an effect, often unintentionally. In the context of AI system deployment, reinforcing feedback loops can occur when the output of a biased model is used to retrain the model, leading to the amplification of existing biases (Weerts, 2019). This cycle can cause the bias to become more entrenched over time, reducing the fairness and effectiveness of the AI system and potentially leading to increasingly skewed outcomes.

As highlighted in the 'Verification and Validation' phase, automation bias refers to the tendency to favor suggestions or decisions made by AI systems, even when there are warning signals or conflicting information from other sources (Khera et al., 2023).

Recommendations for addressing reinforcing feedback loops and automation bias

- **Implement continuous monitoring and documentation**: Document how the performance of the model is monitored and ensure that the system and its feedback loops are continuously observed to detect and address any emerging biases or patterns of reinforcement (Nazer et al., 2023). Regular monitoring is crucial to identifying potential issues early and preventing biases from becoming entrenched.
- **Incorporate community-driven data and neutral labels**: Incorporate community-driven data and add 'neutral' labels to new data. This approach can reduce the risk of biases being reinforced as the model evolves (Muhammad, 2022).
- **Monitor for data drift**: While a model may perform well initially after deployment, its performance can degrade over time due to data drift—changes in the feature distribution of the data received in production. This shift can cause the model's output to become biased, and if this biased output is used to retrain the model, it can lead to a reinforcing feedback loop. Stakeholders should actively monitor for signs of data drift and take corrective actions as needed to maintain the model's performance and fairness (Nazer et al., 2023).
- **Human oversight and critical evaluation**: Stakeholders should remain vigilant to the possibility of automation bias emerging, even if it is not evident initially. Over time, this bias can start to influence decision-making processes subtly. If signs of automation bias do appear, it is crucial to address them promptly and ensure that human oversight and critical evaluation remain integral parts of the AI system's operation (Khera et al., 2023).

Example: Reinforcing Feedback Loops in Predictive Policing

Law enforcement agencies often use predictive policing algorithms to allocate resources based on crime data. However, if these algorithms disproportionately target certain communities due to historical biases in crime reporting or enforcement, it can result in increased surveillance and policing in those areas. This, in turn, can further reinforce negative stereotypes and biases, creating a reinforcing feedback loop where biased data leads to biased outcomes, which then perpetuate the initial biases in the system.

PHASE 6 CONTINUOUS VALIDATION

TYPES OF BIAS IN THIS PHASE

- Data biases
- Evaluation bias

'Continuous Validation' is a crucial phase in the lifecycle of AI systems that employ continuous learning but 'it is also applicable in other situations without continuous learning, for example, to detect data drift, concept drift or to detect any technical malfunctions' (ISO/IEC 5338). In systems with continuous learning, models integrate new data on an ongoing basis without explicit retraining. This dynamic use of new data necessitates regular checks to ensure that the new data aligns with the original dataset and that the model's performance remains up to standard. Additionally, the test data itself may need periodic updates to better reflect the current deployment environment and ensure its relevance.

During this phase, stakeholders consistently assess the AI system's performance using updated test data to verify correct operation before returning to the 'Operation and Monitoring' phase. The success and effectiveness of AI systems with continuous learning heavily depend on the quality of their implementation. If models are poorly integrated or if their quality deteriorates significantly when new data is incorporated, the system will struggle to adapt to its environment and may fail to perform as intended (Pianykh et al., 2020).

DATA BIASES

In continuous learning, where new data is integrated into the model without explicit retraining, it becomes essential to thoroughly assess the new data for potential biases. If the new data introduced during the learning process contains biases, the model may perpetuate or even amplify these biases over time.

Several types of data-related biases, previously discussed in the 'Data Understanding and Preparation' phase, are particularly relevant in this context:

- **Representation Bias**: Occurs when the data does not accurately reflect the diversity of the population, leading to skewed model outcomes.
- **Selection Bias**: Arises when the data collection or sampling methods result in a non-representative sample, which can mislead the model's predictions.
- **Measurement Bias**: Involves systematic errors in data collection that disproportionately affect certain groups, leading to inaccurate predictions.
- **Historical Bias**: Reflects the inherent biases present in the historical data, which can cause the model to replicate past injustices.
- **Labeling Bias**: Occurs when inconsistencies in labeling lead to biased training data, impacting the model's performance.
- **Proxies**: When indirect measures are used as substitutes for difficult-to-measure variables, they can unintentionally introduce bias if they correlate with sensitive attributes.

Incorporating new data in continuous learning requires careful evaluation to ensure that these biases are not introduced or exacerbated, maintaining the fairness and accuracy of the AI system over time.

Recommendations for addressing data bias

Continuous learning relies on the critical assumption that the new data being fed into the system is of high quality and representative of the underlying distribution. However, if the incoming data contains errors, noise, or systematic biases, it can significantly degrade the model's performance and reliability (Pianykh et al., 2020).

To ensure the model continues to perform effectively, stakeholders must be vigilant in monitoring and mitigating potential biases as they arise during this phase. By following the data-related recommendations mentioned earlier, such as those addressing representation, selection, and measurement biases, stakeholders can help maintain the integrity and fairness of the AI system throughout its continuous learning process.

EVALUATION BIAS

As previously discussed in the 'Validation and Verification' phase, evaluation bias occurs when the metrics and procedures used to assess the model's performance are not appropriately aligned with the model, dataset, or the target population. This misalignment can lead to inaccurate evaluations, as the chosen metrics may not accurately represent how the model will perform in real-world scenarios (Suresh & Guttag, 2021). This bias is especially critical in continuous learning systems, where ongoing evaluation is necessary to ensure that the model adapts effectively to new data without compromising its fairness and accuracy.

Recommendations for addressing evaluation bias

- **Constantly question model performance**: Continuous learning models require ongoing scrutiny to ensure that their performance remains robust over time. It is essential to regularly evaluate the relevance and appropriateness of the evaluation metrics being used (Pianykh et al., 2020). This practice helps ensure that the model's outputs are still aligned with its intended objectives and that it continues to perform well across diverse scenarios.
- **Ensure proper oversight and governance**: Without proper oversight and governance, continuous learning systems may inadvertently adopt unethical or harmful behaviors from the data they process. This could lead to unintended consequences, such as privacy violations, the spread of misinformation, or the reinforcement of harmful stereotypes. To prevent these outcomes, stakeholders must implement strong governance frameworks that include regular audits, ethical guidelines, and checks to ensure that the model's learning processes are aligned with ethical standards and societal values.

Example: Continuous Learning in Chatbots

Chatbots deployed in customer service applications often use continuous learning to enhance their ability to understand and respond to user queries more accurately over time. As new conversational data becomes available, the chatbot adapts its responses based on user feedback and refines its language understanding capabilities without requiring manual retraining. This continuous learning process allows the chatbot to stay up-to-date with evolving language patterns and user preferences, thereby improving its overall performance and effectiveness in real-world scenarios.

However, without proper oversight, the chatbot could misinterpret user feedback or responses, leading to inaccurate or inappropriate replies. This misinterpretation can result in user frustration and dissatisfaction with the chatbot's performance, ultimately leading to a negative user experience and potential damage to the company's reputation. Ensuring continuous monitoring and governance of the learning process is crucial to maintaining the chatbot's effectiveness and preventing unintended consequences.

PHASE 7

RE-EVALUATION

TYPES OF BIAS IN THIS PHASE

- Evaluation bias
- Abstraction traps
- Sunk-cost fallacy
- Status Quo bias

The 'Re-evaluation' phase follows the 'Operation and Monitoring' phase and is a critical stage in the life cycle of an AI system. During this phase, stakeholders conduct a thorough assessment of the system's performance, comparing it against the initial objectives and identified risks. This evaluation serves as a pivotal moment for refining objectives and requirements based on the insights gained from the system's operational experience. If the evaluation reveals that certain metrics or outcomes are inadequate, the AI system may need to revisit earlier stages, such as the 'Inception' or 'Design' and 'Development' phases. This allows for the refinement of objectives or the exploration of different approaches to address the problem more effectively. The 'Re-evaluation' phase ensures that the AI system remains aligned with its goals and can adapt to changing requirements or challenges.

EVALUATION BIAS & ABSTRACTION TRAPS

In the 'Re-evaluation' phase, the performance of the AI system is carefully assessed and compared with the original objectives, while also identifying any new or ongoing risks. During this phase, it is crucial to be mindful of potential evaluation bias and abstraction traps, which were discussed earlier.

Evaluation bias can distort the assessment of the system's performance if the metrics and procedures used are not appropriately aligned with the model's intended use or the population it serves. Abstraction traps, on the other hand, may cause stakeholders to overlook important contextual factors by oversimplifying the problem or the system's interaction with its environment.

Awareness of these biases and traps is essential for ensuring that the re-evaluation process provides a comprehensive and accurate reflection of the AI system's effectiveness and fairness. By addressing these potential pitfalls, stakeholders can make well-informed decisions about whether to refine the system or explore alternative approaches.

SUNK-COST FALLACY

The Sunk-Cost Fallacy is the tendency for individuals or organizations to continue investing in a project or endeavor in which they have already invested significant resources—such as money, time, and effort—even when the current costs outweigh the potential benefits (Haita-Falah, 2017).

In the context of AI systems, stakeholders might be influenced by the substantial resources they have invested in developing and deploying the system. This can lead to a reluctance to abandon or significantly alter the system, resulting in a continuous effort to fix issues that may not be worth the cost or might even be impossible to resolve. The larger the sunk cost, the stronger the presence of this bias (Haita-Falah, 2017).

This bias can lead to the deployment of AI systems beyond their optimal lifespan, potentially causing harm if the system begins to operate unfairly or ineffectively. It is crucial for stakeholders to recognize the sunk-cost fallacy and make decisions based on the current and future value of the system, rather than past investments.

Recommendations for addressing sunk-cost fallacy

- **Involve diverse perspectives and expertise**: Engaging a diverse group of stakeholders in the re-evaluation process can help counteract individual biases, including the sunk-cost fallacy. By incorporating varied perspectives and expertise, the decision-making process becomes more balanced and rational, reducing the likelihood of being swayed by past investments.
- **Establish clear evaluation criteria**: Setting clear, objective criteria for evaluating the system's effectiveness is crucial. These criteria should be closely aligned with the overarching goals of the project. This approach helps guide stakeholders in making informed decisions based on the system's current and future value, rather than being influenced by the resources already invested. By focusing on these well-defined metrics, the influence of sunk costs can be minimized, leading to more objective and rational outcomes.

STATUS QUO BIAS

Status Quo Bias refers to the preference for maintaining the current state of affairs, leading to resistance to change (Gong, 2015). In the context of evaluating an AI system, stakeholders may unconsciously exhibit this bias, resisting necessary changes to the system's operation even when such changes are crucial for improving its effectiveness and fairness. This reluctance to alter the existing system can hinder progress and prevent the implementation of improvements that are needed to ensure the system continues to meet its objectives in a dynamic environment.

Recommendations for addressing status quo bias

- **Increase awareness and evaluate the AI system**: As with any cognitive bias, increasing awareness is the first step in avoiding the Status Quo Bias. It is crucial to take the time to thoroughly evaluate the AI system and carefully weigh all possible options for adaptations (The Decision Lab, n.d.-a).
- **Implement team evaluations and collaborative discussions**: To effectively counteract this bias, it is advisable to conduct the evaluation collectively as a team rather than relying on individual assessments. Collaborative discussions and comprehensive consideration allow the team to pool diverse perspectives, leading to a more well-informed evaluation. This collective approach helps ensure that necessary changes are recognized and implemented, and it fosters the development of a strategic plan that aligns with the system's long-term goals and effectiveness.

Example: Addressing Status Quo Bias in Traffic Optimization AI

An AI system deployed to optimize traffic flow in cities fails to meet its initial objectives of reducing congestion and improving commute times, primarily due to inaccurate predictions of traffic patterns. Despite these shortcomings, stakeholders recognize the need for a thorough assessment rather than clinging to the current system. After careful evaluation, the system is returned to the 'Inception' phase, where objectives are refined, and data collection methods are enhanced. This proactive approach helps overcome the Status Quo Bias, allowing the stakeholders to make necessary changes that improve the system's effectiveness and better align it with its goals.

PHASE **8**

RETIREMENT

TYPES OF BIAS IN THIS PHASE

- Historical bias (legacy bias)

As AI systems evolve and adapt to changing circumstances, there may come a point when they no longer effectively meet evolving needs and requirements. This marks the beginning of the 'Retirement' phase, where stakeholders must make critical decisions about the future of the AI system. Retirement may be necessary if the system is no longer required, a superior alternative emerges, or if the system produces unfair outcomes that cannot be rectified in earlier phases. A loss of trust in the system due to these issues may also drive the decision to retire it.

HISTORICAL BIAS (LEGACY BIAS)

During the 'Retirement' phase, stakeholders need to consider the potential lingering effects of the AI system, even after it has been decommissioned. Without proper governance measures, underlying data, processes, or decisions from a retired AI system could continue to influence future systems. For example, data from a retired, biased system might still be used in new models, perpetuating the same biases in different forms. Addressing these residual impacts is crucial to prevent ongoing harm.

The previously mentioned Child Benefit Scandal in the Netherlands serves as a stark example of how biased AI systems can contribute to long-lasting and damaging consequences. Despite efforts to address the issue, the effects of this scandal persist years later, impacting both the financial and psychological well-being of those affected.

This case highlights how the consequences of biased AI systems can linger in society long after the systems themselves have been retired. Therefore, it is crucial for stakeholders to carefully manage the 'Retirement' phase, ensuring that any lasting effects are mitigated and that lessons are learned to prevent similar issues in the future.

Conclusions

In this document, we have explored the various biases that can emerge throughout the life cycle of an AI system. As we've seen, addressing biases in AI cannot be achieved solely through technical solutions. Beyond statistical biases, AI systems are also vulnerable to biases stemming from stereotypes, inherent prejudices, and cognitive biases that may unconsciously influence human judgment and decision-making.

It is important to recognize that biases will inevitably be present in AI systems, and the pursuit of completely unbiased systems is unattainable. Instead, our focus should be on understanding the nature and impact of these biases and determining what level of bias is acceptable within the specific context of each AI system. Biases are often context-dependent, and there is no one-size-fits-all solution. Therefore, individuals and organizations must remain vigilant and proactive throughout the entire AI life cycle.

Stakeholders must collaborate to incorporate diverse perspectives and expertise, particularly from domain experts, as the responsibility of mitigating bias should not fall on a select few. It is crucial to consistently ask critical questions about the data, algorithms, and decision-making processes involved, and to challenge assumptions and biases at every stage. To this end, creating and using diverse and representative datasets, developing and implementing rigorous testing and validation protocols, and conducting ongoing monitoring and evaluation of model performance are essential practices. Bias mitigation and risk assessment tools can significantly support these efforts.

This document serves as an initial overview for individuals and stakeholders navigating the complexities of bias in AI systems. While it may not cover every possible bias, it provides a foundation for further exploration, with additional resources provided below for continued learning and action.

Additional Resources

To read more about bias, fairness, and discrimination, you can check out the following resources:

- 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), 1–35. https://doi.org/10.1145/3457607
- Muhammad, S. (2022, May 11). The Fairness Handbook. Retrieved from https://openresearch.amsterdam/en/page/87589/the-fairness-handbook
- Schwartz, R., Vassilev, A., Greene, K., Perine, L., Burt, A., & Hall, P. B. (2022, March). Towards a standard for identifying and managing bias in artificial intelligence. https://doi.org/10.6028/nist.sp.1270
- Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and Abstraction in Sociotechnical Systems. Association for Computing Machinery. https://doi.org/10.1145/3287560.3287598
- Suresh, H., & Guttag, J. V. (2021). A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. AMC. https://doi.org/10.1145/3465416.3483305
- UK Information Commissioner's Office: https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/guidance-on-ai-and-data-protection/annex-a-fairness-in-the-ai-lifecycle/
- (Only in Dutch) Ministerie van Binnenlandse Zaken en Koninkrijksrelaties. (2022, January 18). Handreiking non-discriminatie by design. Retrieved from https://www.rijksoverheid.nl/documenten/rapporten/2021/06/10/handreiking-non-discriminatie-by-design

Additionally, if you're looking to implement fairness metrics, these open-source toolkits can assist in assessing and mitigating bias and related fairness issues throughout the AI lifecycle:

- Fairlearn: https://fairlearn.org/
- AI Fairness 360: https://aif360.res.ibm.com/
- FairTest: https://fairtest.org/
- What-If Tool: https://pair-code.github.io/what-if-tool/
- TensorFlow Fairness Indicators: https://www.tensorflow.org/tfx/guide/fairness_indicators

FINAL OVERVIEW

AI Lifecycle Phase	Potential Bias	Recommendations
1. Inception	Institutional Bias	• Identify and involve diverse stakeholders • Define affected demographic groups
	Abstraction Traps: The Formalism Trap	• Align problem formulation with social context
	Abstraction Traps: The Ripple Effect Trap	• Consider the broader impact on all actors • Analyze power dynamics
	Abstraction Traps: The Solutionism Trap	• Critically evaluate the need for AI
2.1. Design	Construct Validity Bias	• Collect multiple measures for complex constructs • Document and report considerations for target variables and features • Acknowledge variability in interpretation of features • Begin considering which fairness definitions and metrics will be applied, and why these choices are appropriate for the specific AI system being developed
	Abstraction Traps: The Framing Trap	• Implement problem framing and solution evaluation
	Abstraction Traps: The Portability Trap	• Consider contextual differences
2.2 Data Understanding and Preparation	Representation Bias	• Ensure balanced representation • Apply sampling techniques
	Selection Bias	• Ensure balanced representation • Apply sampling techniques
	Measurement Bias	• Re-evaluate the measurement process • Collaborate with domain experts
	Labeling Bias	• Define clear labeling requirements • Determine labeling methods and performance metrics
	Historical Bias	• Improve representation of minority groups • Collaborate with domain experts
	Proxies	• Select appropriate proxies • Consider sensitive features
2.3. Development	Algorithmic Bias	• Employ de-biasing techniques • Ensure transparency, interpretability, and reproducibility • Incorporate fairness constraints
	Aggregation Bias	• Incorporate group differences into the objective function • Address underfitting
	Omitted-variable Bias	• Apply feature importance methods
	Learning Bias	• Critically select optimization metrics • Address representation bias
3.Verification and Validation	Evaluation Bias	• Assess the suitability of evaluation metrics • Compare performance across groups • Mitigate overfitting • Monitor data distribution imbalances
	Automation Bias	• Promote awareness and training
	Anchoring Bias	• Promote awareness and training
4. Deployment	Deployment Bias	• Promote stakeholder discussions • Enhance model interpretability and understandability
	Abstraction Traps	• Follow the recommendations for these types of bias
	Popularity Bias	• Promote awareness and training • Implement technical methods
5. Operation and Monitoring	Reinforcing Feedback Loop	• Implement continuous monitoring and documentation • Incorporate community-driven data and neutral labels • Monitor for data drift
	Automation Bias	• Implement human oversight and critical evaluation
6. Continuous Validation	Data Biases	• Follow data-related recommendations to address representation, selection, and measurement biases
	Evaluation Bias	• Constantly question model performance • Ensure proper oversight and governance
7. Re-evaluation	Evaluation Bias and Abstraction Traps	• Follow the recommendations for these types of bias
	Sunk-Cost Fallacy	• Involve diverse perspectives and expertise • Establish clear evaluation criteria
	Status Quo Bias	• Increase awareness • Implement team evaluations and collaborative discussions
8. Retirement	Historical bias (legacy bias)	• Implement governance measures to address residual impact

References

Amnesty, (2021). Xenophobic machines Dutch child benefit scandal. Retrieved from https://www.amnesty.org/en/latest/news/2021/10/xenophobic-machines-dutch-child-benefit-scandal/

Barocas, S., & Selbst, A. D. (2016). Big Data's Disparate Impact. California Law Review, 104(3), 671. https://doi.org/10.15779/z38bg31

Borgesius, F. Z. (2018, January). Discrimination, artificial intelligence, and algorithmic decision-making. Strasbourg: Council of Europe, Directorate General of Democracy. Retrieved from https://pure.uva.nl/ws/files/42473478/32226549.pdf

Equinet (2021). Expanding the list of protected grounds within anti-discrimination law in the EU. Retrieved from https://equineteurope.org/expanding-the-list-of-protected-grounds-within-anti-discrimination-law-in-the-eu-an-equinet-report/

European Union Agency for Fundamental Rights. (2022). Bias in algorithms – Artificial intelligence and discrimination. Publications Office of the European Union. Retrieved from https://data.europa.eu/doi/10.2811/25847

Geng, S. (2015). Decision Time, Consideration Time, And Status Quo Bias. Economic Inquiry, 54(1), 433–449. https://doi.org/10.1111/ecin.12239

Gichoya, J. W., Thomas, K., Celi, L. A., Safdar, N. M., Banerjee, I., Banja, J. D., . . . Purkayastha, S. (2023). AI pitfalls and what not to do: Mitigating bias in AI. British Journal of Radiology, 96(1150). https://doi.org/10.1259/bjr.20230023

Haita-Falah, C. (2017). Sunk-cost fallacy and cognitive ability in individual decision-making. Journal of Economic Psychology, 58, 44–59. https://doi.org/10.1016/j.joep.2016.12.001

Haselton, M. G., Nettle, D., & Andrews, P. W. (2015). The Evolution of Cognitive Bias (pp. 724–746). https://doi.org/10.1002/9780470939376.ch25

ISO/IEC 22989:2022. (2022). Retrieved from https://www.iso.org/standard/74296.html

ISO/IEC TR 24027:2021. (2021). Retrieved from https://www.iso.org/standard/77607.html

Jacobs, A. Z., & Wallach, H. (2021). Measurement and Fairness. Association for Computing Machinery. https://doi.org/10.1145/3442188.3445901

Jiang, H., & Nachum, O. (2019). Identifying and correcting label bias in machine learning. arXiv (Cornell University). https://doi.org/10.48550/arxiv.1901.04966

Khera, R., Simon, M. A., & Ross, J. S. (2023). Automation bias and assistive AI. JAMA, 330(23), 2255. https://doi.org/10.1001/jama.2023.22557

Klimashevskaia et al. (2024). A Survey on Popularity Bias in Recommender Systems. https://arxiv.org/abs/2308.01118

Kordzadeh, N., & Ghasemaghaei, M. (2021). Algorithmic bias: review, synthesis, and future research directions. European Journal of Information Systems, 31(3), 388–409. https://doi.org/10.1080/0960085x.2021.1927212

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), 1–35. https://doi.org/10.1145/3457607

Ministerie van Binnenlandse Zaken en Koninkrijksrelaties. (2022, January 18). Handreiking non-discriminatie by design. Retrieved from https://www.rijksoverheid.nl/documenten/rapporten/2021/06/10/handreiking-non-discriminatie-by-design Ministerie van Justitie en Veiligheid. (2024, March 26). Wat is discriminatie? Retrieved from https://www.mensenrechten.nl/mensenrechten-voor-jou/discriminatie-en-gelijke-behandeling/wat-is-discriminatie

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

Muhammad, S. (2022, May 11). The Fairness Handbook. Retrieved from https://openresearch.amsterdam/en/page/87589/thefairness-handbook

Naghiaei, M., Rahmani, H. A., & Dehghan, M. (2022). The unfairness of popularity bias in book recommendation. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2202.13446

Nazer, L., Zatarah, R., Waldrip, S., Ke, J. X. C., Moukheiber, M., Khanna, A. K., . . . Mathur, P. (2023). Bias in artificial intelligence algorithms and recommendations for mitigation. PLOS Digital Health, 2(6), e0000278. https://doi.org/10.1371/journal.pdig.0000278

Nickerson, R. S. (1998). Confirmation bias: a ubiquitous phenomenon in many guises. Review of General Psychology, 2(2), 175–220. https://doi.org/10.1037/1089-2680.2.2.175

Ntoutsi, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdl, W., Vidal, M., . . . Staab, S. (2020). Bias in data-driven artificial intelligence systems—An introductory survey. WIREs Data Mining and Knowledge Discovery, 10(3). https://doi.org/10.1002/widm.1356

Pianykh, O. S., Langs, G., Dewey, M., Enzmann, D. R., Herold, C., Schoenberg, S. O., & Brink, J. A. (2020). Continuous Learning AI in radiology: implementation principles and early applications. Radiology, 297(1), 6–14. https://doi.org/10.1148/radiol.2020200038

Rastogi, C., Zhang, Y., Wei, D., Varshney, K. R., Dhurandhar, A., & Tomsett, R. (2022). Deciding fast and slow: The role of cognitive biases in AI-assisted decision-making. Proceedings of the ACM on Human-computer Interaction, 6(CSCW1), 1–22. https://doi.org/10.1145/3512930

Ruf, B., & Detyniecki, M. (2021). Towards the right kind of fairness in AI. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2102.08453

Santiago, T. (2019). AI bias: How does AI influence the executive function of business leaders? Muma Business Review, 3, 181–192. https://doi.org/10.28945/4380

Schwartz, R., Vassilev, A., Greene, K., Perine, L., Burt, A., & Hall, P. B. (2022, March). Towards a standard for identifying and managing bias in artificial intelligence. https://doi.org/10.6028/nist.sp.1270

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and Abstraction in Sociotechnical Systems. Association for Computing Machinery. https://doi.org/10.1145/3287560.3287598

Suresh, H., & Guttag, J. V. (2021). A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. AMC. https://doi.org/10.1145/3465416.3483305

The Decision Lab. (n.d.-a). Status quo bias. Retrieved from https://thedecisionlab.com/biases/status-quo-bias

The Decision Lab. (n.d.-b). The sunk cost fallacy. Retrieved from https://thedecisionlab.com/biases/the-sunk-cost-fallacy

Ulnicane, I., & Aden, A. S. (2023). Power and politics in framing bias in Artificial Intelligence policy. Review of Policy Research, 40(5), 665–687. https://doi.org/10.1111/ropr.12567

Van Giffen, B., Herhausen, D., & Fahse, T. (2022). Overcoming the pitfalls and perils of algorithms: A classification of machine learning biases and mitigation methods. Journal of Business Research, 144, 93–106. https://doi.org/10.1016/j.jbusres.2022.01.076

Varona D. and Suárez J.L. Discrimination, bias, fairness, and trustworthy ai. Applied Sciences, 12(12):5826, Jun 2022. https://doi.org/10.3390/app12125826

Weerts, H. (2021). An introduction to algorithmic fairness. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2105.05595

Wet AwGB. (2020, January 1). Retrieved from https://wetten.overheid.nl/BWBR0006502/2020-01-01

Žliobaitė, I., & Custers, B. (2016). Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models. Artificial Intelligence and Law, 24(2), 183–201. https://doi.org/10.1007/s10506-016-9182-5

Acknowledgements

We would like to extend our sincere gratitude to the following individuals and organizations who contributed to the development of this research:

Radboud University in Nijmegen, The Netherlands
We deeply appreciate the collaboration with the Radboud University, particularly with the student from the MSc in Artificial Intelligence Susanne Snoek and her supervisor Haselager W.F.G. We want to thank Suzanne for her dedication to this research. Her analysis has been crucial to form the basis for this study.

Student Researcher and Author:
Suzanne Snoek

Rhite Consultants
Special thanks to the team at Rhite for guiding the direction and focus of this research.

Lead Researcher and Supervisor:
Isabel Barberá

Support Team:
Lea Magnano (designer)
Alexia Muresan (reviewer)

Your collective efforts and commitment to advancing the field of Trustworthy AI have been invaluable to this project!

Contacts

For further discussion or questions regarding this research, please contact:
Head of AI Research at Rhite: **Isabel Barberá**
Email: **info@rhite.tech**

General Inquiries: **Rhite**
Website: **www.rhite.tech**
Email: **info@rhite.tech**

We welcome your feedback and look forward to engaging with you on the important topic of bias detection and mitigation in AI systems.

PRACTICAL GUIDE

From Inception to Retirement: Addressing Bias Throughout the Lifecycle of AI Systems

Leading the way to Trustworthy AI

Visit our website

www.rhite.tech

Rhite

Leading the way to Trustworthy AI