AI Fairness-by-Design Multi-Stakeholder Methodology

-

A Comprehensive Framework for Fair AI Design and Development

AEQUITAS
unbias AI

ANNEX V: Fair Model Methodology (FMM)

Fair-by-Design sub-methodology

Abbreviation	Meaning
AFF	Affectees
AIU	AI Users
DDM	Development Decisionmakers
DE	Domain Experts
EGTAI	Ethics Guidelines for Trustworthy AI
FbD	Fair-By-Design
FDCGM	Fair Data Collection, Governance, and Management
FMM	Fair Model Methodology
FOIM	Fair Output Interpretation Methodology
FRIA-F	Fundamental Rights Impact Assessment for Fairness
GDM	Governance Decisionmakers
SIM	Stakeholder Identification Methodology
TAIRA	Trustworthy AI Readiness Assessment
MAP	Multistakeholder Approach to AI Fairness-by-Design
ML	Machine learning
NLP	Natural Language Processing

AEQUITAS
unbias AI

Contents

Introduction

The Fair Model Methodology (FMM) will assist Development Decisionmakers (DDM) during the development stage of the AI system to select or design a fair model in light of the intended purpose (envisioned solution) at hand. In this deliverable, we make the difference between DDM who designs a model from scratch and DDM who selects an existing model (with or without modifications). The same principles will have to be considered by both groups, the only difference is that the first one will follow the proposed processes when designing the model, while the second group will first follow the requirements when selecting the model and then follow the process in case iterations are made on the algorithm.

There are multiple sub-fields of AI and training techniques that this Deliverable aims to cover, to provide a comprehensive methodology. As a starting point, it is paramount to acknowledge that there is no universally accepted definition of an AI system or AI model. Guidance can be found in numerous textbooks, guidelines and self-regulatory instruments such as standards, but from a normative perspective, it is important to identify the definition of an AI system of the EU AI Act.

We focus on how algorithms are designed, developed, and trained, how they generate outputs from the provided inputs, and which outputs/roles the training aims for.

The main aim of this Deliverable is to provide a methodology on:

- How to select the type of algorithm/model most suitable for the intended purpose at hand, taking into account the relevant context, including the relevant ethical, legal and social constraints of that context;
- How to design a fair AI model, including specifying if and how legal, ethical, and social constraints should be addressed by the technical components and processes of building the model that enables the AI system to achieve the objectives it is designed for.

For DDM, this step will come in continuity with what was shared in the Fair Data Collection, Governance, and Management Methodology (FDCGM) and it will be initiated at the development stage.

From the literature review shared in **Deliverable 5.2** (P.2), there is a multitude of studies and techniques enabling fairness in Machine Learning (ML) and Deep Learning (DL) considering the exponential use of these AI techniques. However, the other techniques and their specificities to enable fairness fall short with a small number of solutions and academic interest observed from an analysis of the state of the art. Therefore, **in this document we will try to address a larger range of AI subfields to fill the existing gap, make the methodology applicable to a wider range of cases, and assess the way the legal, social, and ethical constraints will shift based on the specificities of each technique**.

The Specificities of Modelling and their Implications on Fairness

Modelling is a scientific technique originating in mathematics and statistics, that has expanded first in the fields of **hard science**. Nevertheless, the idea that human thoughts are merely a

combination of calculations has arisen from the philosophical work of Hobbes and Leibniz, and Turing and Church worked on the idea of translating 'mental calculations' to machines[1].

At the early stages of the AI field, there were already ethical and sociological studies on the potential challenges of model-based automation of thought processes and there is still a lack of full understanding of all variables that should be considered when designing algorithms for processes outside hard science. Hence, particularly in the fields of social sciences, law and ethics, rules cannot simply be translated into mathematical equations. This **historical background** illustrates the first challenge in enabling fairness when designing AI systems in contexts where mere inferences from mathematical equations can have serious implications on people's rights, ethical principles and social values.

AI models are also often taken as the end state of a process (even representing the "truth"), rather than the beginning of a discussion, that includes proper assessment and nuance. Marc Jacobs & Ronald Meester, in their book "From earthquake to zoonosis – On the use of models for policy" argue that "by objectifying models and seeing them as "truth" rather than a helpful representation to better understand dynamics of a complex, wicked issue, we have stretched their use into dogmatic reasoning." They believe this represents a danger to our society and democracy and highlight the necessity for a "**mathematical model to be the beginning of a dialogue, not the (definitive) end**". This re-emphasises the importance of considering the field, use case, and context of the model's deployment for the algorithm to serve its purpose.

This will require the fair collection, governance and management of training data, which is something that was addressed in the Fair Data Collection, Governance, and Management Methodology (FDCM).

During the design or selection of the algorithm, the **modeler makes assumptions** about the dataset which, along with the dataset's features, shape the model. Therefore, it is paramount for DDM, who directly build the model, to be aware of the legal, social, and ethical constraints that their assumptions should incorporate. This is important to enable the fairness of the model and consequently its robustness, accuracy, and applicability. On one hand, modelers need to be aware that **the features of the dataset and the related assumptions** can **hold ethical, legal, or social meaning** which is beyond the technical skills expected from DDM. Hence the relevance of the **interdisciplinary working group** proposed in the Stakeholder Engagement Methodology which will collaboratively identify, assess, and manage the legal, technical, social, and ethical risks of the AI system. On the other hand, modelers need to be aware of **the impact the model will have** and the use case where it will be implemented to build a relevant model fair and applicable in the context purposed. These insights will be gathered from **Domain Experts (DE), Affectees (AFF), and AI Users (AIU) by the Governance Decisionmakers (GDM)** and coordinated with the DDMs in the Scoping and Risk Analysis Stages.

AI systems are not only used to describe and understand phenomena, but also to **predict events and prescribe solutions**. These operations are executed in several domains and fields such as banking, recruitment, and policymaking where complex issues are in hand. For

[1] Turing Alan, Computing machinery and intelligence (London: Mind, vol. 59, no 236, 1950) 433-460.

instance, during the COVID-19 pandemic crisis, policymakers governed the pandemic period using statistical predictions of contamination leading to lockdowns.

AI Model Characteristics Challenging Fairness

AI-driven processes have a set of characteristics which differentiate between AI systems and other new technologies, making both its specificity and challenge. Scherer, M. provided a relevant analysis of the special characteristics of AI from where risk can emanate. He provides a risk typology of AI characteristics which make regulating the technology challenging for decision-makers namely; autonomy, foreseeability, causation, discreetness, discreteness, diffuseness, and opacity.[2]

Autonomy, Foreseeability, and Causation (ex-post)

AI Systems can execute complex tasks autonomously making their outputs non-predictable by humans in situations where the model creates new correlations (e.g.: AlphaGo). In several use cases, such as building a portfolio or driving airplanes, AI systems are currently able to accomplish a set of actions without the intervention of humans. The autonomous characteristic of AI is specific to this technology making it possible to free humans from multiple tasks, raising efficiency, and optimizing profits. However, this delegation of tasks from humans to machines conceals decisions that hold, in some cases, ethical, legal, or social values. Therefore, it is paramount to align the already established socio-ethical and regulatory principles to the outputs of the system. To ensure fairness in the Model's outputs, related considerations should be included already in its design.

AI systems have a higher computational capacity compared to humans considering the large amount of data they can cover and their technical objectivity when creating correlations. These characteristics enable the machine to show new outputs different from the ones provided by human analysis. As humans tend to be opinionated, subconsciously biased, or subscribing to conventional wisdom, modelling provides a pure calculation to explain events and take action. This is exactly what makes AI systems efficient, objective, and sometimes a fairer solution. Nevertheless, **not all objective correlations between variables of a dataset are ethically, legally, and socially fair**. Therefore, we cannot count on the Model to provide fair outputs without designing it for fair principles in advance and monitoring it for non-foreseeable risks.

The question of control becomes more and more relevant with the increase of AI systems' capabilities. The loss of control over the AI system does not imply the scenario of an out-of-control big robot killing humans, but it is much more concrete. A cybersecurity breach, flawed programming, or corrupted file in the dataset are all illustrations of how a loss of control over an AI system[3] may occur, enabling the machine to autonomously achieve unintended understandings, actions and purposes to ensure **as much control as possible** in building a fair AI system by eliminating/mitigating the risks of bias at the design stage of the model. Keeping **humans in the loop** and developing mechanisms for intervention in case of unfairness incidents are paramount to maintaining control over the AI system.

[2] Scherer, M. U. (2015, May 30). Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies. *Harvard Journal of Law & Technology.* Volume 29, Number 2. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2609777
[3] Scherer, M. U. (2015, May 30). Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies. *Harvard Journal of Law & Technology.* Volume 29, Number 2. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2609777

Discreetness, discreteness, diffuseness, and opacity

The discreetness, discreteness, diffuseness, and opacity principles are related to the way AI research and development processes are done. Scherer, M. highlights that these principles are not specific to AI systems compared to other technologies, contrary to the outlined ones in the section above. However, they are argued to present unique challenges in this technology[4].

In fact, the infrastructure behind AI systems is often discrete with limited visibility over the dataset and a lack of understanding of the model's correlations. Therefore, WP6 proposes throughout its deliverables to **endorse transparency**, document all steps of the AI lifecycle stages, and consult with GDM, DE, AIU, and AFF which makes the process gain visibility and external oversight.

Diffuseness and **discreteness** of AI systems are closely related; one refers to the way an AI component can be developed by many individuals in different places, and the second considers the way different components of AI systems are developed across diverse timelines and locations without coordination. Hence the relevance of WP6's approach in this methodology specifying **model design and selection processes**, and consequently including prerogatives on which information should be shared by the AI developer on their risk management strategy.

The opacity of AI systems is illustrated when the model cannot be revealed, or reverse engineered. Despite the rise of open-source initiatives, modelling takes large resources which necessitate developers to generate a return on investment for contributors by profiting from their proprietary of the model. Consequently, the explainability of the model can become further challenging in these contexts. Therefore, it is important for DDM to incorporate **explainability enablers** when designing the model to understand the outcomes including when they are unfair. This does not only help in understanding the reason of the issue but also simplifies solving it which provides **better paths for remedy and redress when incidents occur**.

Assessing and enabling fairness at the model level

Since there is a multitude of variables to consider before building a fair model, interdisciplinary collaboration is paramount when designing the algorithm. To simplify the process, WP6 provides a set of questions that GDM and DDM should discuss to cover all these dimensions (partially inspired by the eleven questions conceived by Marc Jacobs & Ronald Meester in the context of policymaking).

[4] Scherer, M. U. (2015, May 30). Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies. *Harvard Journal of Law & Technology*. Volume 29, Number 2. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2609777

As mentioned before, DDMs are not expected to have a legal, ethical, nor social background. Therefore, WP6 will provide in this section the principles, processes, and steps DDMs need to follow when designing, selecting, or modifying an algorithm.

Fairness considerations regarding the type of model

AI systems can have different roles, depending on the purpose their outcomes serve. We make the difference between predictive, descriptive, and prescriptive models where fairness considerations change based on the technical specificities of each. To achieve a comprehensive fair design, it is paramount to not hinder the technical characteristics implied by the role of the model along with the legal, ethical, and social constraints.

Descriptive Models

Descriptive models, as their name indicates, are designed to describe, explain, and understand phenomena based on the provided dataset. In this case, the algorithm is designed to find correlations between the dataset variables using statistical analysis techniques. There are several techniques employed to enhance the diagnosis of data and its control, namely, Expert Systems, Fuzzy Logic, and Pattern Recognition[5]. These techniques differ in their design and in the way their outputs are provided, but they all serve to find relationships between the variables, providing structural, behavioural, or other descriptions.

Making sense of a large amount of data raises efficiency in understanding previous and present elements by reducing time for humans, increasing capacity, and enabling a better-informed analysis and decision. Therefore, descriptive models have been increasingly implemented across diverse fields. For instance, descriptive models assist researchers in understanding further their scope of research and inform them of what the empirical data holds about the topic. Inside business organizations, marketing, sales, and finance teams, among others, descriptive models are implemented to improve their understanding and decision-making based on insights from previous and present data.

It goes without saying that at this stage the dataset should already be assessed for biases by mitigating legal, ethical, social, and technical sources of unfairness. The focus must be on how to design a fair descriptive model considering the previously identified risks and the technical specificities of this type of AI system. The algorithm encapsulates interconnections among components that signify its requirements, architecture, behaviour, and parametric limitations. DDMs must make sure that these interconnections are not unfair to any group of people, entities, or components of the environment. This can be done by adding rules on which labelled variables shouldn't be correlated, for instance.

Due to its modelling language accommodating multiple abstraction methods, the system model also facilitates the depiction of diverse perspectives of the system, including black-box, white-box, and security viewpoints. Additionally, the system model can be examined and queried for coherence, functioning as a unifying framework. [6]

[5] G, Krithiga, V. Mohan, and S. Senthilkumar. "A Brief Review of the Development Path of Artificial Intelligence and its Subfields." International Journal of Engineering Technologies and Management Research 10, no. 6 (June 3, 2023): 1–12. https://doi.org/10.29121/ijetmr.v10.i6.2023.1331.
[6] Friedenthal , S. et al. (2024). "A Practical Guide to SysML: The Systems Modeling Language." (Chapter 2). The MK/OMG Press.

Predictive Models

Predictive Models are designed to predict future events based on historical data. Depending on the context where it is deployed, the algorithm can predict upcoming sales, protein combinations for antibodies, or criminal recidivism of a suspect. There are diverse techniques used in designing predictive models, such as Artificial Neural Networks and Support Vector Machine[7]. A Neural Network is a model that creates patterns by mimicking the way human brains operate. Composed of input and output layers, each artificial neuron is connected to other ones and has an associated weight and threshold that contribute to the algorithmic conclusions.[8]

The impact of predictive algorithms varies based on the context where they are deployed with some high-risk applications that require risk management during the design of the model. If the designed model will, for instance, be deliberating judicial-related decisions, it will have a direct say in co-deciding the innocence of suspects. Therefore, it is paramount for DDMs to study the context of deployment before the design of the model considering:

- The existing rules and laws in the domain to dodge irrelevant inferences. One process that can help in distinguishing between **hard and soft science**. Inference processes in hard science combines objective data while soft science contexts are more complex due to the personal and sensitive data contributing to outcomes. In the latter, it is paramount to make sure that inferences are not created based on name, ethnicity, socio-economic class, or gender, among other demographic data that can be grounds for discrimination.
- If the data is labelled, **supervised learning** is preferred, over unsupervised, to control the way the model learns and make sure bias is mitigated.
- The **environmental impact of training**, especially that large models, tend to require high amounts of electricity which emits tons of carbon emissions[9]. DDMs are encouraged to use designed and trained models.

It is worth noting that the aim is not to lose the technical performance of predictive models which have shown to be transformative across different fields in the last years. But it is important to ensure that the model is fair, explainable, and controllable so the algorithmic predictions that decision-makers consider are legitimate, accurate, and understandable.

Prescriptive Models

Prescriptive Models are designed to recommend actions to solve the problem at hand based on the analysed data. The models do not only suggest guidance based on existing patterns but also predict the potential outcomes of each solution to benchmark the most efficient one.[10]

[7] G, Krithiga, V. Mohan, and S. Senthilkumar. “A Brief Review of the Development Path of Artificial Intelligence and its Subfields.” International Journal of Engineering Technologies and Management Research 10, no. 6 (June 3, 2023): 1–12. https://doi.org/10.29121/ijetmr.v10.i6.2023.1331.
[8] IBM. “What Are Neural Networks?” Www.ibm.com, IBM, 2023, www.ibm.com/topics/neural-networks.
[9] Ren, Shaolei, and Adam Wierman. “The Uneven Distribution of AI’s Environmental Impacts.” Harvard Business Review, Harvard Business Publishing, 15 July 2024, hbr.org/2024/07/the-uneven-distribution-of-ais-environmental-impacts.
[10] Bazzarelli, Manuela. “Predictive, Descriptive and Prescriptive Models.” Aramix, 10 July 2023, aramix.ai/en/blog/ai-models/predictive-descriptive-and-prescriptive-models/.

There are diverse techniques used to design prescriptive models such as Intelligence Agents and Stochastic Techniques.[11]

The level of risk of prescriptive models also depends on the context as they can prescribe the most competitive pricing strategy or suggest militarily targeting a house based on the existence of a potential terrorist. Therefore, DDMs will have to consider, on top of the provided suggestions for a fair predictive model, the limitations of the model in decision-making, by incentivising to use the outcomes as the start of the discussion, not the final decision.

Fairness considerations during modelling

Vigilance in AI modelling is crucial to address bias, in addition to the fairness mechanisms adopted in the training and use of the system. As explained in Deliverable 5.1, "bias in modelling may, bias 'in modelling ' can be deliberately introduced, for instance, through smoothing or regularization parameters to mitigate or compensate for bias in the data (algorithmic processing bias), or when using objective categories to make subjective judgments (algorithmic focus bias)." Bias can emanate from diverse sources as illustrated by Suresh & Guttag's taxonomy illustrated in the figure below.[12]

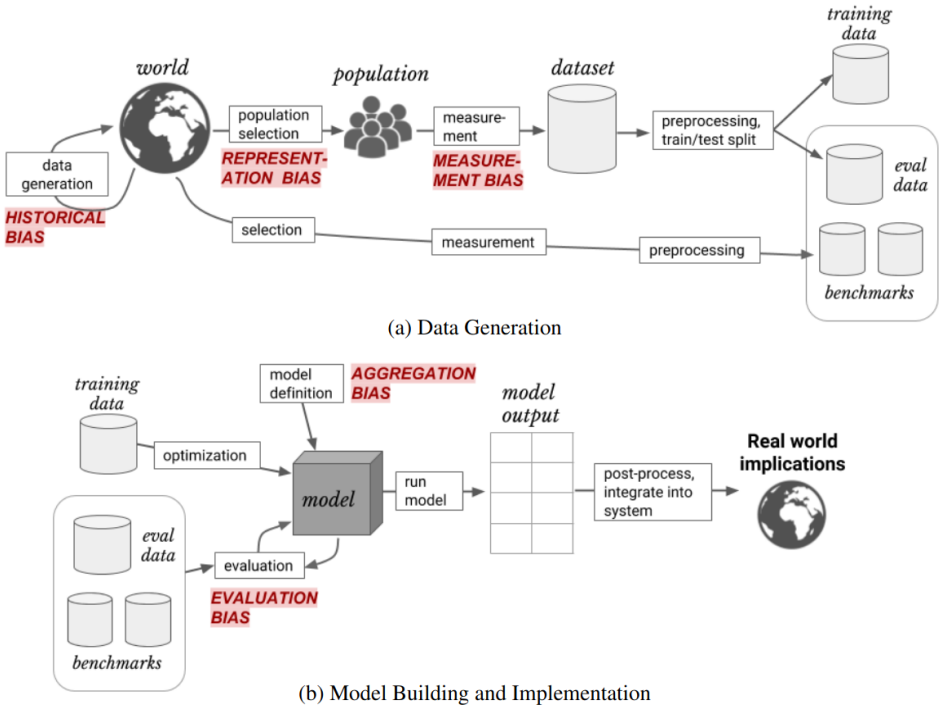


Figure 1: Bias taxonomy following the framework of Suresh and Guttag (2019).

According to the steps of the development and implementation of the AI system, the figure above illustrates how risks of bias may be introduced. Starting from historical bias existing in

[11] G, Krithiga, V. Mohan, and S. Senthilkumar. "A Brief Review of the Development Path of Artificial Intelligence and its Subfields." International Journal of Engineering Technologies and Management Research 10, no. 6 (June 3, 2023): 1–12. https://doi.org/10.29121/ijetmr.v10.i6.2023.1331.

[12] Suresh, H., Guttag, J. 2019. "A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle." Equity and Access in Algorithms, Mechanisms, and Optimization, October. https://doi.org/10.1145/3465416.3483305.

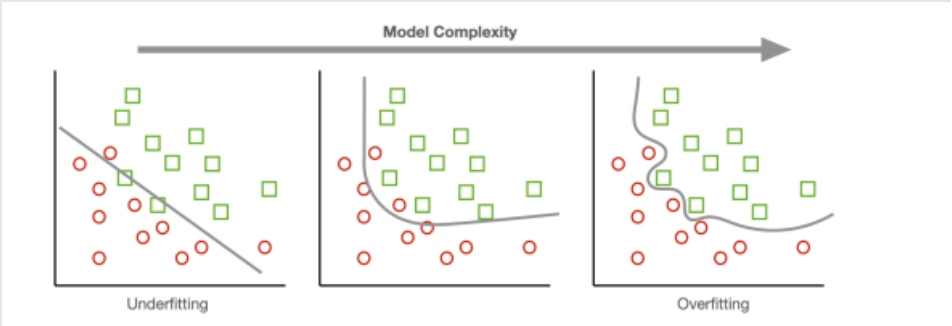
the world to representation and measurement bias which are relevant at the data governance level where GDM and DDM will follow the Fair Data Collection, Management, Governance Methodology in enabling a fair governance of data. Concerning the model definition, **aggregation bias** may arise when correlations and inferences are made inappropriately, discriminating against a certain group. When it comes to the evaluation and interpretation of the model's outcomes, fairness steps for these processes are to be underlined in the evaluation section in the next deliverable and the Fair Interpretation Methodology.

Since the focus of this document is the design and selection of a fair model, we will explain further in the following sections the technical constraints to fairness an algorithm may lead to. The challenges to watch out for are Overfitting/Underfitting; Trade-Offs; Design Specification; Model Hallucination; and Model Shifts (Distribution & Domain Shifts).

Model overfitting & underfitting:

Model overfitting occurs when a model **learns the training data too well**, grasping and learning the noise and details that are not relevant to the general underlying pattern. This algorithmic phenomenon leads to a **model that performs well on the training data but poorly on the newly introduced one**. One of the causes of model overfitting is using a model that is **too complex for the task at hand**. For instance, using a deep neural network with many layers on a task that does not use large datasets. It is important for DDM to choose a model with **capabilities that are harmonized with the volume of the data available and the nature of the task**. The capacity of the model is related to its expressive power; "low-capacity models will underfit, while high-capacity models can overfit [13]." To control the complexity of the model, DDM can implement **regularization processes** which vary based on the type of the model. For example, DDM can control Decision Trees by restricting the number of nodes in the tree, while in Neural Networks, the magnitude of the weights can be determined to control the model and minimize the occurrence of overfitting.

Addressing overfitting in ML still represents a challenge; 152 studies using predictive ML models were evaluated in terms of their methodological quality revealing that only half of them evaluated the model for overfitting with accurate strategies [14]. Overfitting is indeed a common challenge in the validation of the model, but it would be efficient to implement predictive measures before the evaluation of the model already during the design of the algorithm to minimize the risk.



[13] Cunningham. P., Delany. S. 2021. "Underestimation Bias and Underfitting in Machine Learning." https://arxiv.org/pdf/2005.09052
[14] Navarro C.L.A., Damen J.A.A., et al. 2021. "Risk of bias in studies on prediction models developed using supervised machine learning techniques: systematic review." BMJ. 2021;375:n2281. doi: 10.1136/bmj.n2281

Figure 2 : Model Complexity, overfitting, and underfitting [15]

The figure above showcases model overfitting and underfitting in relation to two classes of the data and their relation to model complexity. While overfitted models processes and learns from noise, underfitted models fails to consider important details.[16]

Model underfitting happens when a model is **too simple to capture the underlying patterns in the data**. In this case, the algorithm fails to learn from the training data and **performs inadequately on both training and newly introduced data**. This phenomenon can occur if the model is **too simple for the task**. For instance, when a linear model is designed for a complex problem falling short due to its limited capacity to comprehensively learn patterns in the dataset.

Model underfitting can also emerge if the algorithm **is not trained for a sufficient amount of time**, reducing the learned correlations from the data. Finally, if the DDM makes **incorrect assumptions**, the model may not align with the underlying structure of the data. For instance, if a linear relation was assumed by DDM while the actual relationship is non-linear. Therefore, it is paramount for DDM to keep in mind the nature of their decisions when designing the model making sure that **no unfair assumptions** are made. It is also important to **monitor if a classifier underfits variables in the dataset** ensuring a fair estimation of variables to dodge biases. The classifier could under predict positive outcomes for minority groups, leading to a systemic disadvantage.

Fairness issues when dealing with trade-offs:

When eliminating or mitigating the risk of unfairness in the model, **DDM might make trade-offs**, shifting the priorities and calculations of the algorithm. For instance, if the DDM tweaks the model to treat all variables/AFF equally, **implications on accuracy for specific groups may occur**[17]. Not all biases are discriminatory, depending on the context and task at hand, DDM must **pay close attention to the correlations that define the task** (e.g., previous fraud leading to non-allocation of credit) and **the ones that lead to unfairness**. The **classifier used to enable fairness** can be the exact reason for the accuracy decrease, knowing that this risk mitigation action constrains the model's solution to the provided problem[18]. It is also worth mentioning that modelling can initially decrease bias by eliminating experts' subjective perspectives and allowing the study of large amounts of data. The need to maintain the benefits of the pure calculation of existing correlations while eliminating unfairness makes balancing trade-offs more challenging.

The early-stage **choices of optimization and performance metrics** are crucial in addressing trade-offs since the design of the model. However, equalizing both accuracy/calibration and

[15] Cunningham. P., Delany. S. 2021. "Underestimation Bias and Underfitting in Machine Learning." https://arxiv.org/pdf/2005.09052

[16] Cunningham. P., Delany. S. 2021. "Underestimation Bias and Underfitting in Machine Learning." https://arxiv.org/pdf/2005.09052

[17] Ferrara, E. 2023. "Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies." Sci 6 (1): 3. https://doi.org/10.3390/sci6010003.

[18] Pessach, D., and Erez, S. 2023. "A Review on Fairness in Machine Learning." ACM Computing Surveys 55 (3): 1–44. https://doi.org/10.1145/3494672.

odds is argued to be incompatible, as illustrated in the COMPAS case study.[19] Pessach, D., et al. (2022) advice to **choose one of these goals** in accordance with the case study at hand.

Model Shifts:

A **model shift** (drift) occurs when the algorithmic performance degrades gradually, and it can take different forms, such as **distribution or domain shifts**. If, between the training and deployment phases, **variations in the statistical characteristics of the data** occur, the algorithm will be faced with solving inferences in a dataset with new distributions. **Distribution shifts** in the model could lead to unfair outputs due to the **lack of training on new data**, especially when it relates to demographics or has any socio-economic impact on individuals. To dodge such a phenomenon, DDM can **include metrics that minimize model bias**, such as **demographic parity**, to incorporate the changing properties of data distribution in the deployment phase.

In **practical implementations** and **dynamic societal applications** of AI systems across a long period, the assumption that data distribution stays stationary during deployment does not stand[20]. Despite the efforts to enable fairness in the model's design, the assumption of stationary environments also does not stand in real-world implementations, leading, in many cases, to unfair outputs[21].

Domain shift occurs when the **algorithm is used in a context or domain different** from the training one. This includes **variations in the user base**, **devices of implementation**, and the **environment**. If the model does not incorporate context flexibility during the training phase, its outputs could carry biases and unfair components when the model is faced with a different context. For instance, if the algorithm is trained in urban developed areas, it could be inaccurate when deployed in rural, under-developed areas.

DDM must **find the appropriate tools to sustain the model's accuracy across distribution and context shifts**. For instance, in **ML**, Schumann, C. et al. constructed a **general approach to domain adaptation for fairness** "that covers a wide variety of fairness challenges, from proxies of sensitive attributes, to applying models in unanticipated settings. Within this general formulation, they have provided theoretical bounds on the transfer of fairness for equal opportunity and equalized odds using both VC-dimension and Rademacher Complexity. Based on this theory, they developed a new modelling approach to transfer fairness to a given target domain[22]"

Documenting the process in a Model Card

Model cards are a form of documentation designed to track the technical features of the algorithm in simplified language. They enable further transparency, and accountability, and provide details for users simplifying the maintenance process[23]. Proposed in 2018, model cards do not follow a unanimous format which also depends on the context. There are however

[19] Pessach, D., et al. (2022). A Review on Fairness in Machine Learning. ACM Computing Surveys, 55(3), 1–44. https://doi.org/10.1145/3494672
[20] Stan, S., & Rostami, M. (2023). Preserving Fairness in AI under Domain Shift. ArXiv.org. https://arxiv.org/abs/2301.12369
[21] An, B., Che, Z., Ding, M., & Huang, F. (2022). Transferring Fairness under Distribution Shifts via Fair Consistency Regularization. ArXiv.org. https://arxiv.org/abs/2206.12796
[22] Schumann, C., et al. (2019). Transfer of Machine Learning Fairness across Domains. https://arxiv.org/pdf/1906.09688
[23] Mitchell et al. Model Cards for Model Reporting. ArXiv, Published 5 Oct 2018. Revised 14 Jan. 2019. https://arxiv.org/pdf/1810.03993

some commonly reported features describing the model, its architecture, intended uses, performance metrics and benchmarks, risk management (e.g.: bias mitigation techniques), and limitations.

The Components of the Model Card should be considered by DDMs during the design of the model to make sure that all elements are ready for documentation in the end of the training.

Questions	**Considerations**
What is the additional value of an algorithmic solution compared to the traditional processes? What are their known strengths and weaknesses?	This should be discussed within the specific solution the model provides and the use case/context where the AI system will be applied. It can also build on the results of question zero and detail the discussion in terms of technical performance.
In which field will (was, in case the team selects a designed model) the model be trained and applied; (hard or soft sciences)?	If it is applied to soft sciences, or hard sciences with a direct impact on Humans, extra attention should be given to the variables and correlations of the model considering the legal, ethical, and social constraints.
Where will the AI system be used and which impact would it have on the modelling requirements?	To answer question 3, the following three questions should be addressed by considering the results of the risk analysis stage and identifying risks of unfairness. Which affectees will be impacted by the system and how should the algorithm be designed to mitigate known and foreseeable risks of unfairness? In which domains will the system be used and what are their implications on the model design (In terms of existing laws and inferences in the discipline/domain)? Who will be using the system, which level of technical knowledge does this group have, and how can the design of the model accommodate knowledge gaps or other shortfalls (if any)? How does this system fit into existing workflows and what organizational processes will need to change?
What is the role/type of the model: prescriptive, descriptive, or predictive?	**In Descriptive Models,** the algorithm encapsulates interconnections among components that signify its requirements, architecture, behaviour, and parametric limitations. Make sure that these interconnections are not unfair to any group of people, entities, or components of the environment. **Predictive Models** are designed to predict future events based on historical data. Consider the context of deployment before the design/selection of the model considering: • A critical evaluation of what is to be predicted (human behaviour, sales expectations, defects of malfunctions, etc.)

	• A critical evaluation of whether making predictions is safe and justified given the context of use (rights to essential public or private services, functioning or management of critical infrastructure, law enforcement, migration, asylum, judiciary, democratic processes, etc.) • A critical evaluation of whether the future events of interest can reasonably be predicted? • Existing rules and laws in the context of use. • If the data is labelled, supervised learning is preferred, over unsupervised, to control the way the model learns and make sure bias is mitigated. • The environmental impact of training, especially that large models, tend to require high amounts of electricity which emits tons of carbon emissions. DDMs are encouraged to use existing designed and trained models. **Prescriptive Models** are designed to recommend actions to solve the problem at hand based on the analysed data. The models do not only suggest guidance based on existing patterns but also predict the potential outcomes of each solution to benchmark the most efficient one. Consider the same recommendations established for predictive models. Additionally, they need to consider the limitations of the model's performance and the AI system's decision-making, by incentivizing the use of the outcomes as the start of the discussion, not the final decision.
Which limitations does the model hold in the prescription, description, or prediction of a phenomenon?	Based on the identified characteristics of the model's role (descriptive, prescriptive, or predictive), identify its potential risk to fairness.
Can, and if so how will, risks to fairness be addressed (or how were they addressed, if the model was acquired?	Identify technical solutions to address these risks fairness when designing the model. Or, if the model is externally acquired, the technical solutions that were implemented to address these risks.
Which techniques will be (were, if the model is selected) used for modelling and what is their impact on the model's fairness?	Example: NLP, CV, ML, Fuzzy Logic...
Which assumptions should the modeler establish in the design of the model? If the model is selected, add: who constructed the model, and what interests, if any, do the creators themselves have in it?	A list of potential assumptions to be made (or already made, in case the model is selected) Do these assumptions hold legal, ethical, or social dimensions? Which risks of unfairness are observed in making these assumptions? Which assumptions can be fair, considering the identified risks?
Which potential misuses could be observed at this stage? And, if the model is	Detail a list of intended & prohibited uses, and potential misuses.

selected, does the provider list any use framework?	
How could potential misuses be mitigated at the modelling stage? And/Or, if the model is selected, does the provider list any mitigated misuses?	Focus on technical solutions or/and disclaimers to add at a later stage in the Model Card.