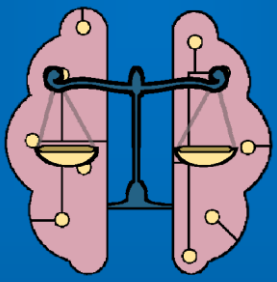
The AI Creator's AI Fairness Definition Guide



Version 1

February 2024

Emmanouil Krasanakis, Marta Gibin, Stavroula Rizou,
and the MAMMOth consortium

This research work was funded by the European Union under the Horizon Europe MAMMOth project, Grant Agreement ID: 101070285. UK participant in Horizon Europe Project MAMMOth is supported by UKRI grant number 10041914 (Trilateral Research LTD). Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or European Commission. Neither the European Union nor the granting authority can be held responsible for them. The content of this document is © of the authors.

Abstract

The adoption of Artificial Intelligence (AI) in all sorts of automated systems and digital services creates concerns over biases with real world ramifications and harm. This guide aspires to help system creators (e.g., engineers working on the research, design, implementation, and deployment of AI models and user interfaces) understand how to define fairness in the social context of their systems by working together with stakeholders and experts from other disciplines. It does so by presenting a workflow that transcribes abstract fairness concerns of affected stakeholders to corresponding formalisms and practices that combine computer, social, and legal science viewpoints.

Table of Contents

Introduction

Artificial Intelligence (AI) systems that incorporate computer science algorithms and techniques to imitate human reasoning affect people's lives by automating predictions, scoring, and recommendations. There are increasing concerns of whether these systems are fair, especially since they tend to replicate or exacerbate real-world biases, or find spurious correlations between certain demographics and predictions.

The first step in diagnosing and treating biased AI consists of concretely defining fairness. This assists the creation of policies and regulations, prevents the deployment of harmful systems, and supports algorithmic solutions for bias mitigation. However, the meaning of fairness depends on who is asked, and on the regulations and stakeholders that systems encounter in different markets. It also varies across predictive tasks (e.g., financial services, face verification, news recommendation, scientific search). Thus, it is hard to create one operational definition that caters to all AI systems, stakeholders, and contexts.

The lack of a one-size-fits-all definition of fairness is reflected in the many formalisms (e.g., mathematical measures or constraints) and practices presented in this document, which capture different opinions and contexts, and often contradict each other (Kleinberg, 2016). For example, proportional representation of minorities in positive classification outcomes could be at odds with equal accuracy between the minorities and the entire population (see mathematical definitions).

As a system creator, you are called to either select definitions of fairness from algorithmic fairness literature or devise new ones. Your systems require technical definitions (e.g., measures, constraints, or software engineering practices) to develop and evaluate, but social and legal sciences often express fairness in abstract terms, such as protecting

underprivileged groups or participating in societal improvement. These terms organize the opinions of individuals into schools of thought or policies that differ from technical language (e.g., they may be hard to quantify). This guide will help you orchestrate an interdisciplinary collaboration that converts non-technical concepts into technical ones.

A multidisciplinary approach to fairness

The authors worked on bridging the gap between abstract fairness statements and concrete definitions within the MAMMOth European Union (EU) project. This tackles bias under the intersection of multiple sensitive attributes (e.g., gender, race) and data source modalities (e.g., tabular data, images, graphs), and one of its goals is to operationalize definitions of bias while accepting input from affected stakeholders. The project created a unique opportunity for merging the understanding of fairness across computer, social, and legal sciences into one workflow that accounts for context-dependent concerns. We followed this to extract MAMMOth's software and research requirements.

This guide publicizes the above-mentioned workflow, so that AI system creators can use it to derive fairness definitions tailored to their operational environment, either by selecting them from the growing literature or by implementing new ones. We structure the proposed process as a series of five questions, which you can answer on a per-case basis to arrive at sensible definitions for your systems. These are:

Who should define fairness?

Why is fairness needed?

Where & When should fairness be applied?

What is considered a fair outcome?

Which practices should be followed?

The first two questions help extract the social context, and the last three identify how to incorporate context-specific concerns in real-world systems. Final fairness definitions are not limited to philosophical terms or mathematical formalism, like probabilistic constraints, but also include fair practices to be followed.

Challenges

We reconcile complex sociological perspectives with software engineering by viewing AI systems as socio-technical ones, which means that they comprise intertwined technical and social aspects. For such systems, we summarize the challenges we mentioned so far in the following three statements.

Fairness is context-specific.

There is no general fairness definition that applies to every context or use case. In this guide, you will find –among other things– an overview of popular definitions from computer science literature. However, which ones are suitable depends on the specific situation you are studying; less common defisnitions could be preferable in certain cases.

There can be conflicting interests and opinions on what is fair.

Different stakeholders might have different ideas on what constitutes a fair solution to a problem. Think of an example AI system that evaluates loan requests: bank clients might want their personal circumstances to be part of the evaluation, but lenders might think it is fair to provide impartial and systematic responses (although these may also contain biases that were not accounted for during system creation, like historical racism in training data).

To make matters worse, there may be power relations and conflicting interests between stakeholders that require negotiation of interests and opinions. This guide distinguishes between this negotiation, which is delegated to social science processes, and implementing opinions that are eventually selected as important.

Fairness is multi-layered.

That is, it needs to account for various aspects, such as technical, social, legal, and ethical. This guide is meant for AI system creators, so there is a focus on the technical aspects. However, these make up only a part of the problem; throughout this guide we recommend close cooperation with other disciplines to properly address the issue of fairness.

How this guide helps

While selecting or creating fairness definitions for your systems from a wide range of feasible options, the presented methodology will help you acknowledge the views of affected stakeholders by working with social scientists and comply with the regulatory backbone of AI markets by working with legal experts.

To this end, you will first get a clear idea of involved parties and their roles (Who), and be introduced to social science fairness objectives (Why). You will also find out where fairness interventions may occur within an AI system's pipeline and lifecycle (Where & When), and become familiar with common types of fairness definitions that may be reused or created, be they formalism (What) or scientific and legal practices (Which).

Parts of the process we outline are already followed by many works, including research papers that start from specific concerns and contexts, propose suitable definitions of fairness, and then satisfy them through

algorithmic means. This guide structures how you should systematically approach the creation of similar fairness-aware AI systems, without getting lost amidst the multitude of new fairness initiatives and regulations. Importantly, it will let you grasp the whole picture by looking at fairness as a dynamic concept that emerges through the context-dependent friction of different members of society and lies beyond the confines of technical analysis.

Who

should define fairness?

Goal
Learn about parties that should be involved in defining fairness and their roles.

Summary
To create real-life technological solutions, combine research principles with fairness concerns of affected stakeholders. This requires co-designing AI systems with said stakeholders. Consult with legal experts to ensure compliance with laws and regulations, and work with social scientists to gather interests of stakeholders and ensure that they are adequately represented and integrated.

In this guide, we urge that definitions of fairness applied on AI systems should adhere to scientific and legal standards while reflecting the opinions of affected stakeholders, such as business owners, system users, or those indirectly affected by predictions. This constitutes a mixture of *principlism*, which is the practice of following predetermined principles, with *pragmatism*, which refers to acknowledging the reality that stakeholders face. In this section we discuss these standalone approaches and the need of combining them.

Our mixed approach lets you, as a system creator, obtain diverse perspectives from stakeholders through a well-structured process. Indicatively of how important such feedback is (we later present explicit arguments in that front), a broad range of interested parties was also consulted via a survey for the creation of the *Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems* (G7, 2023b), with which this guide is compatible.

As scientific standards are part of system implementation, you are responsible for properly following them. Input from legal experts is needed to follow AI market regulations; your organization may already have a legal team or department that you can work with. Finally, we later explain that it may be difficult to identify and extract opinions from stakeholders; these tasks should be carried out by social scientists (e.g., anthropologists, economists, sociologists, psychologists). For instance, these may identify and work with affected underrepresented groups (e.g., oppressed groups, racial minorities), or organizations that represent them, such as non-profit ones, and organize discussions between multiple stakeholders to reconcile conflicting opinions.[1]

Overall, the combined process we outline involves the following actors: a) system creators, b) affected stakeholders, c) social scientists, and d) legal experts. The combined effort of these is what will define fairness for created AI systems, and selecting specific people or organizations for each role answers the question of who is defining fairness.

Principlism: definition by AI creators

Principlism refers to applying universally accepted –and therefore context-independent– ethical principles on a scientific domain. This term was first coined for medical bioethics (Beauchamp and Childress, 2001) but a similar assumption tends to be implicitly followed by AI fairness research and algorithmic fairness code frameworks (e.g., AIF360); both implement predetermined ethical principles that have been mapped to mathematical definitions. Popular definitions of this kind tend to gain approval by affected stakeholders in many social

[1] Product owners, that is, the people or organizations on whose behalf you create AI, may also be stakeholders whose opinions will be gathered. If you are product owner yourself, provide feedback independently of your capacity as a system creator. Let social scientists lead in any discussions between stakeholders.

contexts (Saxena et al., 2019), although it is unknown whether the same approval would manifest in new contexts.

Regulatory frameworks may also present principles that guide or mandate how fairness is defined. For example, in the EU market, fairness is one of the key requirements presented by the *Ethics Guidelines for Trustworthy AI* (AI HLEG, 2019), and one of the AI principles of the *Organization for Economic Cooperation and Development* (OECD, 2023). Compatibility with provisions or official recommendations requires the involvement of legal experts, especially when technical approaches for quantifying bias become loose indicators under legal framing (Xiang and Raji, 2019). Not all fairness principles are backed by law or are commonly agreed upon, which is why this guide adopts a mixed approach of also incorporating stakeholder opinions.

Sticking to principlism during AI system creation is an attractive prospect in that it involves only a few actors. In fact, under this approach, and besides the help of legal experts on determining compliance with market regulations, system creators would be the ones to select applicable principles from the literature.

For example, the exclusion of protected characteristics (e.g., gender, race, national origin) from decision-making in the field of employment is laid down in *Title VII of the Civil Rights Act*[2] of the US, as amended[3]; this rare alignment between legal and technical language makes it clear that you should follow the algorithmic practice of not using these characteristics in AI predictions in the corresponding market. Furthermore, well-studied scientific standards or mathematical definitions of fairness incorporate corresponding principles into

[2] *See:* Civil Rights Act of 1964 (Public Law 88-352).
[3] *See:* Public Law 114-95, enacted December 10, 2015.

reusable code frameworks, which drastically speed up the creation of new systems with well-tested code.

On the other hand, principlism has certain limitations. To begin with, it is too focused on the technical aspects, and may therefore miss the social context or side effects on the world (John-Mathews et al., 2022). To understand this claim, recall that there might be conflicting ideas of what constitutes fairness among different people, in different contexts, and across different perspectives (e.g., technical, social, legal) (Gerards et al., 2023). However, principlism promotes a top-down approach to defining fairness that is driven by AI creators; as technical definitions constitute only one incomplete perspective (Ruggieri et al. 2023), defining fairness unassisted could result in missing the bigger picture.

In a similar vein, principlism creates the illusion that well-studied definitions of fairness, which are easier to compute or optimize, should be applied in every setting. Trying to circumvent this issue with more abstract principles, as regulations often do, rarely yields unambiguous definitions of fairness to implement, and instead requires context-specific exploration, like the one promoted by pragmatism below.

Example

Let us look at the four-fifths rule, which is widely popular in research circles and code frameworks. This stipulates that the ratio of accepted protected subgroup members should at worst be 80% that of the majority. Mathematically, for protected group G and majority G' for binary acceptance decision $\hat{y} \in \{0,1\}$ of individuals x this rule takes the form:

$$\frac{P(\hat{y} = 1, x \in G)}{P(\hat{y} = 1, x \in G')} \leq 0.8$$

Unfortunately, the four-fifths rule does not satisfy the US legal framework of disparate impact in which it is often used; the latter requires significance testing instead (Biddle, 2017; Watkins et al., 2022) and the rule can be used only a “rule of thumb” that verifies the existence of certain intuitive biases if not met.

Another important limitation of principlism is that seemingly technical decisions can have consequences on people's lives in ways that are not anticipated by the chosen principles. For example, binary classifiers that are made to have the same positive rates between males and females – which is a common process for preventing gender disparate impact– automatically forgo any protection of those who do not recognize themselves in one of these genders. In this case, what appears like a technical choice (usage of a binary protected attribute) is actually a political one that discriminates against a group of people.

Pragmatism: accommodate stakeholder input

Practical concerns of stakeholders are acknowledged by a different approach, called *pragmatism*. In this, instead of assuming the existence of globally true ethical principles, ethics are built from the lived experiences, perceptions, opinions, and concerns of stakeholders (John-Mathews et al., 2022). Discussions with multiple parties can support this process and highlight potentially diverging opinions on fairness (Whittaker, 2019). Particular attention should be paid to the struggles that oppressed groups face in the social situation under analysis.

If you followed a pragmatic approach to gather fairness definitions for an AI system, you would need the assistance of disciplines beyond computer science. First, cooperation with legal experts and social scientists would start from identifying affected stakeholders, such as legally protected and systemically oppressed groups of people in the relevant socio-technical context. Multistakeholder discussions (i.e., discussions involving multiple stakeholders) should then be conducted

to analyze stakeholders' opinions, interests and concerns, and extract definitions of fairness, such as in free text form.[4]

Discussions would be organized and guided by social scientists to focus on the issues of all stakeholders, and to not extend beyond the context at hand, as this may inadvertently port definitions to settings they were not meant to address. You and legal experts would also participate, to respectively ensure that extracted definitions of fairness fit in your algorithms, and that regulations are followed.

> Example
> As an example of pragmatic feedback, consider a security-critical facial recognition system deployed in a geographic region. Technical expertise may dictate that it is harder to analyze the faces of racial minorities, due to underrepresentation in training data. Therefore, to safeguard a fair use of the system for racial minorities, representative individuals from those groups, together with social scientists studying the issues they are facing, should be involved in determining which system property to focus on. One kind of feedback could be not to mistake same-minority members for each other (Cavazos, 2020). Different concerns may arise for facial recognition in different settings.

The main barriers in pursuing pragmatism are available resource limits, such as the time, budget, and effort to identify stakeholders and engage them in meaningful discussions. Furthermore, not everyone has the necessary expertise to evaluate the technical details of an AI application (e.g., members of underrepresented groups should not be expected to directly write specifications or design fiction). Hence, a transcription of such details to layperson terms is also needed.

A second shortcoming of interacting with all actors through pragmatism is that it forces them to work together when in practice they tackle

[4] For example questions that could drive discussions with stakeholders, look at the UnBias toolkit.

different parts of the fairness definition problem, thus creating an inefficient democratization of AI (Himmelreich, 2023). In particular, as a primarily social science approach, pragmatism aims to summarize context-dependent concerns of stakeholders via interaction with social scientists. When system creators and legal experts are added in this collaboration, they bring different objectives (e.g., arriving at actionable definitions), and technical or legal language that is harder to align with the ways stakeholders understand and communicate social issues.

Combining principlism and pragmatism

Principlism, like quantifying fairness with the previously described four-fifths rule, and pragmatism that acknowledges input from stakeholders are *not* mutually exclusive. In fact, we encourage a combination of both practices, as the first reduces complexity and brings rigorousness, such as adherence to legal guidelines or reuse of systems with high degree of technological readiness, and the second acknowledges that systems are being applied on and affect the real world.

We hereby describe an approach that aims to address the limitations of principlism with the incorporation of pragmatic practices. This way, we let stakeholder feedback drive the selection, adjustment, or creation of concrete definitions of fairness (Wachter 2021; Weinberg, 2022). Overall, our approach comprises the following steps:

Preparation

1. System creators seek legal experts and social scientists to work with.
2. Social scientists identify affected stakeholders.
3. System creators perform preliminary analysis based on principlism.

Iterative feedback (repeat until stakeholders are satisfied)

4. System creators provide analysis outputs to social scientists.

5. Social scientists employ sociological methods to transfer this analysis to stakeholders and lead pragmatic discussions.
6. Social scientists gather input from stakeholder discussions, like textual stipulations, and provide it to system creators.
7. System creators update definitions of fairness (e.g., mathematical definitions, procedures) while consulting with legal experts.

Given a first identification of related actors in the first two steps, the process of defining fairness definitions from the point of view of system creators starts by searching for a wide variety of applicable fairness definitions, based on principlism. Early prototyping or dataset auditing may help make this search tractable, and you can generate Mepham matrices (O'Neil, 2020) by imagining yourself in the place of hypothetical stakeholders to identify risks against their wellbeing, autonomy, and justice.

Exploration based on such principles dees not constitute final system evaluation, but rather a way of demonstrating simple concepts or statements about fairness that correspond to a first indication of concerns. For example, you may state that a gender appears on average only one time in the top ten predictions for a recommendation system. Share intuitive concepts like this in easy-to-understand formats (e.g., presentations) with social scientists and affected stakeholders to help them get a sense of algorithmic issues at play.

Then, social scientists can follow a pragmatic approach that lets the stakeholders determine what constitutes fairness using the provided concepts as a basis. Use the newly extracted definitions to update or create new principles and repeat the process until stakeholders are satisfied with the outcome. All the while, enlist the help of legal experts to make sure that you comply with regulations. Make several repetitions of this workflow, including several rounds of stakeholder discussions. These are simpler than a fully pragmatic approach, as each discussion's

goal is not to produce a final definition but to refine existing concerns and compare them with the improved understanding of AI creators.

In the above collaboration, we set social scientists as a buffer between the roles of system creators and stakeholders –including product owners– that are involved in the definition of fairness. We do so because the process is challenging by itself. To give you an idea, consider an AI system that regulates university admissions (this example is retrieved from Costanza-Chock, 2020). We could work on fairness by correcting biases related to several intersecting protected attributes, such as race, gender, disability, or national origin. But working on fairness could also entail deciding whether attribute intersections in the admitted candidates should reproduce the proportions in the general population, or whether the system needs to correct the underadmission of certain groups in previous years. There is no universally right answer, but these aspects should be part of multistakeholder discussions.

The workflow we propose requires collaboration between the actors we already recognized (system creators, social scientists, legal experts, stakeholders) and emphasizes how systems end up affecting stakeholders in practice. Figure 1 summarizes the information flow between the actors; existing technical research is not an actor, but still guides the first system analysis that will be enriched or modified by the pragmatic stakeholder feedback retrieved by social scientists.

The time and effort costs that would have been incurred by adding more actors to the pragmatism approach partially persist, though reduced by clear-cut roles and interactions. Other types of influence between actors, like power dynamics between stakeholders or the influence of laws on everyday life, should be documented during the process of gathering pragmatic feedback.

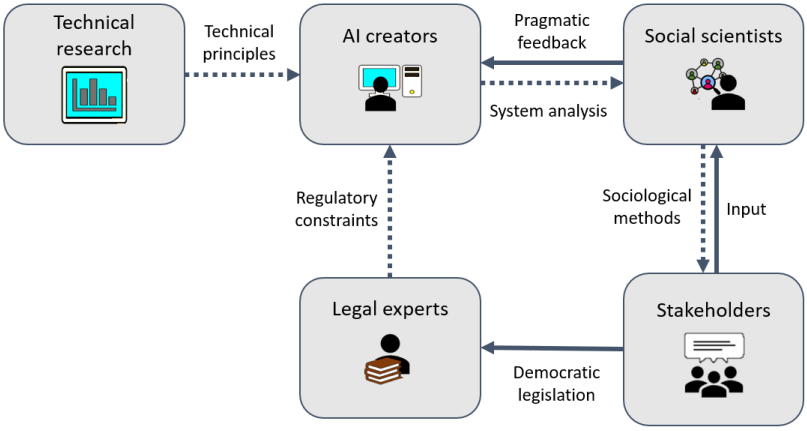


Figure 1. *Information transfer between different actors under our approach for defining AI system fairness; dashed lines transfer principles.*

Why

is fairness needed?

Goal

Learn about the interaction between AI systems and society, and what could be the desired outcome of fairness interventions.

Summary

If left unattended, AI systems may learn from a biased reality and inject back biases that make it even more unfair. Based on stakeholder input and regulation, determine whether you are implementing fairness to passively debias predictions (weak fairness) or to actively participate in societal improvement (strong fairness).

To understand the importance of why any kind of AI system fairness is needed, look at your systems as cogs of society that are prone to creating a feedback loop (Barocas et al., 2017). Without fairness interventions in place, they may not only directly harm individuals, but also exacerbate biases found in the real world, influence reality towards greater unfairness, measure the latter by gathering new or updated data, and then relearn to be even more biased.

This loop is demonstrated in Figure 2 and, if left unattended, may create a vicious cycle of unfairness. However, it also provides an opportunity to improve the world through algorithmic corrections and redesign. For example, instead of exacerbating biases, systems can mitigate or reverse them (e.g., by temporarily creating the opposite biases) and eventually create a more balanced society as their outputs influence reality. Determining the urgency and degree of the sought balance (e.g., accepted trade-offs with system utility) is part of fairness definitions. Thus, it is ideally an input gathered from stakeholders, although it could also be constrained or determined by regulation.

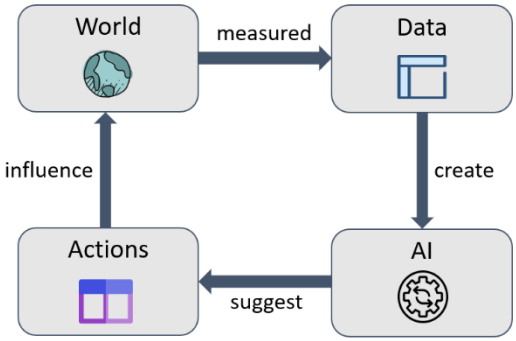


Figure 2. *Feedback between AI and the world – adapted from Barocas et al., 2017.*

Most work on algorithmic fairness aims to achieve what is known as *weak fairness* (Kong, 2022), which debiases AI outcomes without looking at their broader societal impact. However, concerns voiced during multistakeholder discussions, especially by vulnerable groups, might require you to work towards what is known as *strong fairness*, which actively challenges oppression and promotes fairness in society. This section explains the two concepts.

Weak fairness objectives

Weak fairness refers to removing bias from system predictions. Algorithmic bias mitigation research tends to implicitly adopt this goal by making systems reach an approximate equilibrium under some measure of predictive performance when comparing groups of people – usually groups protected by legislation with non-protected ones.

For example, false positive rates (e.g., the rate of wrong predictions for criminal recidivism) may be equalized between different races. Equilibria-based fairness easily leads to mathematical constraints but does not actively challenge oppression. For instance, equal false positive rates between races could give a false semblance of debiasing when

ground truth labels used to compute the rates are also influenced by historical racism of a society.

As an alternative to creating parity between mathematical measures, weak fairness may also look at *individualized symmetrical treatment*, which refers to treating all individuals the same, regardless of the effects of past or present-day discriminations. For instance, this could be expressed as a lack of group bias in the distribution (and not only in an aggregate statistic) of true or false positives, and true or false negatives. What "same treatment" entails depends on how reality is interpreted, and defining it means that input from stakeholders should be incorporated into AI data, for instance through few-shot learning to quantify acceptable predictive differences based on examples.

Although weak fairness does not take drastic action towards improving society, it is not without merit. To begin with, it already breaks the feedback loop of reinforcing bias. For example, preventing a system from replicating racism or sexism found in its historical training data prevents the perpetuation of stereotypes, despite not accounting for the same types of discrimination elsewhere in society. Moreover, it may be challenging to identify a-priori how AI fairness interventions influence predictions or the world, in which case weak fairness becomes the only commonly agreed upon option.

Strong fairness objectives

Strong fairness aims to bring more access, opportunities, and life chances to all people (Kong, 2022). This viewpoint acknowledges that the legacy of long-term discrimination and oppression might require active redistributive actions to challenge discrimination by influencing the real world (Rawls, 1971). In AI terms, oppression can be understood as historical biases that have infiltrated learning data (e.g., biased

jurisdictional decisions), but may also be found in societal structures that correlate with sensitive characteristics (e.g., zip codes correlating with race).

> Example
> Consider, again, the example of university admissions provided in the previous section. Forms of strong fairness could include correcting the underadmission of certain groups in previous years, or placing equal importance on both more and less affordable extracurricular activities that influence the access to universities, given that some groups struggle to pay for expensive ones (Giovanola and Tiribelli, 2022). These should also be determined through stakeholder feedback.

Strong fairness also safeguards every person's *right to justification* (Giovanola and Tiribelli, 2022), that is, the right to receive explanations and understand the reasons behind AI outcomes, and the possibility to appeal decisions they consider unfair. The same prospects have been identified by independent schools of thought too; explainability is a key scientific standard described in the Which section, and the option to appeal decisions is part of *TrustAIOps* that ensure ongoing quality in the Where & When section.

Thinking in terms of strong fairness further means questioning the very purpose of AI systems, what they achieve, and their effects on society. There should be documentation of whose opinions and beliefs are being implemented, who is benefiting from the implementation, and who might be negatively affected. As mentioned before, this requires analysis of the social context and discussions with stakeholders.

Overall, strong fairness reframes the purpose of AI systems to means of improving society. For example, it may make hiring criteria less strict for under-represented genders in certain fields to improve the perception among their members and encourage participation by all genders. Similar practices are known as *inverse discrimination* or *reverse*

discrimination, given that the goal is to actively try to adjust society to protect future individuals at the potential cost of -at first- aiding previously discriminated groups more. Since this can create backlash, you should seek legal advice for the correct interpretation of anti-discrimination mechanisms. Moreover, this approach can cause side effects and unintended consequences, and therefore requires a socio-economic and political impact assessment too.

Where & When

should fairness be applied?

Goal
Get a sense of where fairness definitions can intervene within an AI system lifecycle.

Summary
Making AI systems fair is not an one-time intervention, but a consideration at each step during their lifecycle; fairness spans fair design, development interventions, and ongoing practices to maintain quality.

AI systems are software projects. Thus, their lifecycle spans the three main stages of software development presented in Figure 3: a) design, including planning and requirement gathering, b) development of the main system, and c) monitoring and maintenance. Apart from bug fixing, insights gathered during maintenance may lead to redesign and additional development down the road. During all stages, identify where and when to apply fairness-aware interventions (Calegari et al., 2023; Rana et al., 2023). Interventions include ongoing processes that extract an updated understanding of AI risks and lead to redesign too.

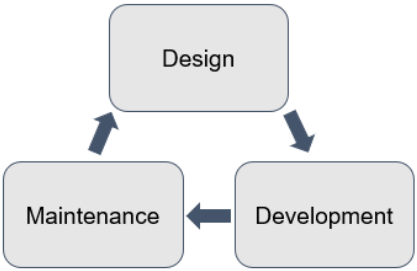


Figure 3. *Software lifecycle.*

Fairness analysis and bias treatment methods could be needed at any step of each one's creation pipeline. To begin with, analysis should start early on, by involving stakeholders in design fiction (e.g., mockups, user stories), and by determining whether business objectives are inherently unjust. For example, widespread population surveillance may be unjust even if it does not discriminate between population subgroups.

During development, identify biases relevant to the selected fairness definitions. Do this for each AI component, before biases are picked up and exacerbated by subsequent components. Do this without losing scientific rigor. For example, do not blindly violate preconditions that algorithms require to properly run, but replace components with equivalent ones that are compatible with fairness interventions.

Some systems, such as chatbots for events, have a short life and are later discarded alongside their results. For these, definitions of fairness like the ones of the What and Which sections are sought mostly during design and development. Systems deployed for longer times also require fairness oversight during maintenance; consider this case the default if you are not sure.

> Example
> To give you an idea of AI systems with especially long lifecycle, they could be multimedia generation or classification services that are accessible through the Internet (e.g., ChatGPT's chatbot, facial identification services), or recommendation components of online platforms (e.g., that select which advertisements to show to each user in search engines or social media).

Maintaining an oversight means that you should not treat technical and non-technical methods for introducing fairness as one-of procedures, but as continuous processes to be followed throughout each system's

lifecycle. We next present in greater detail how fairness concerns can be incorporated into each stage of said lifecycle.

Fair design

There exist various software design methodologies. The *Agile* approach (Beck et al., 2021) quickly iterates through new versions of systems and adjusts them through input from stakeholders, like system owners and users. Other methodologies disentangle design and implementation, thus reducing costly interactions between developers and stakeholders. Regardless of the exact methodology, initial phases include user input[5] and extraction of requirements, usually followed by the construction of design fiction, system architecture, and application interfaces.

When fairness is added as a requirement of AI systems, their design must be modified. For instance, beyond stakeholders that are already involved in the design, such as owners and users, affected parties identified by social scientists and legal experts should also be consulted. Similarly, new system requirements should be extracted, and existing ones adjusted. In practice, it can be hard to involve underrepresented groups in more technical aspects of the design, but you should at least acknowledge their pragmatic expectations.

Some definitions of AI fairness are also imposed "by design" and therefore lead to further modifications at this stage. One such definition is the introduction of counterfactual analysis steps, which require training of corresponding components. Overall, fair system design may require several iterations before implementing even the first

[5] People other than system users may also be affected stakeholders. For example, this may happen if system predictions make the users form opinions or perform actions against certain groups based on past biases.

prototypes. Many of the smaller iterations can be prevented if fairness is recognized early as a system goal.

For longer-life systems, or those designed with Agile methodologies, iterating through new versions requires tracking how bias or fairness considerations evolve. Evolution could arise either from changes in the system itself, changes in the deployment environment (including updates of regulations), or a more advanced understanding of how systems interact with society.

Tracking evolution with regards to fairness has crystallized in the field known as *TrustAIOps* (Li et al., 2023). This follows an Agile methodology that includes iterative refinement of system fairness by gathering feedback during operations in the deployment context. The concept of tracking evolving bias and fairness considerations can be applied to any methodology that relies on iterative refinement of systems.

Fair development

Bias can be unpredictable and arise at different steps of AI system creation, given that different components emphasize different underlying aspects of data. Even worse, subsequent steps may pick up leftover or new unintended biases in the outcomes of previous processing and exacerbate them to an irreversible degree. Understanding where bias comes from should start by scrutinizing your data mining and processing workflow. In some cases, bias could arise from interactions with system creators and their unconscious biases during training and validation, potentially unfair societies from which data are gathered, or biased data gathering practices.

Example

As an example of leftover biases, consider a system for loan approval prediction that employs feature selection as a training data preprocessing step to filter from a long list of available attributes indicating various types of economic activity. If we provide training data that discriminate against women, and which include an attribute (e.g., whether each person has performed a hair transplant) correlated with gender, that attribute is likely to be picked as an important feature in place of other alternatives. Then, the predictor runs the risk of creating sexist decisions based on that attribute and reducing its bias could require revising the feature selection process too.

A typical data mining workflow is presented in Figure 4 alongside example datatypes. In this figure, AI creators work with existing data, preprocess those, and use them to create models, such as neural networks. The outcomes of the latter are post-processed, for example to be converted to formats easily parsable by humans, and finally validated in terms of meeting desired objectives, such as accuracy or fairness. Depending on the outcomes of validation, the steps of this creation process are revisited and modified to address found issues, such as predictive biases.

Asserting fairness cannot be summarized only in terms of one measure extracted at a single step of the above workflow on one dataset; the full process needs to be safeguarded, so that systems can also generalize any lack of bias that is observed in a few (training, validation, and testing) datasets under lab conditions to applications. In other words, AI systems should not only be fair in one circumstance but be *robustly fair*. Conversely, not treating bias at early steps may irrevocably embed it in the dataflow in ways that are hard to catch and quantify later. This last issue will likely be a byproduct of any dimensionality reduction within AI components, for example due to entangling proxies of predictions with sensitive attributes similarly to the previous loan approval example.

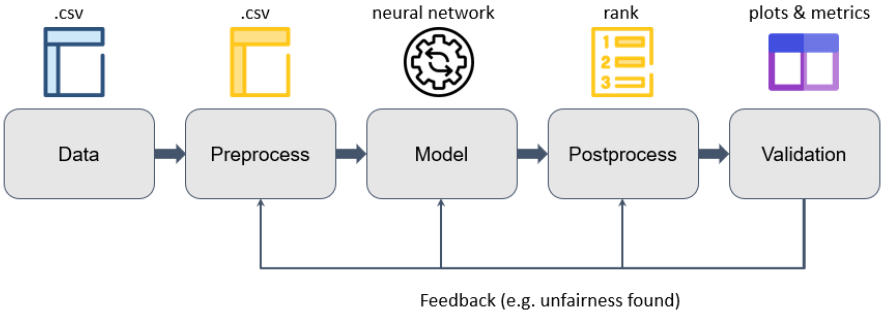


Figure 4. *AI system creation pipeline.*

Ideally, fairness should be assessed just before, during, and after each step. Ask yourself how the outcome of previous steps can affect each subsequent one's inputs, and whether inputs could exhibit biases forewarned by domain expertise, stakeholders, and legal experts. Then, identify mechanisms that work inside steps that may be able to pick up that information and use it to influence outputs. For example, create informed hypotheses on whether AI components identify correlations between protected groups of people, or on whether information is being compressed (e.g., when creating embeddings) with mechanisms that could ignore the behavior of minorities.

Modern AI further tends to reuse pretrained components (especially large language models – LLMs) that serve as embedding mechanisms or publicly available initial states to be fine-tuned on new data. Unfortunately, many publicly available models are known to create biased outputs (e.g., Thakur, 2023 unveils LLM sexism).

If left unchecked, dataset and model biases of reused modules will also leak into your pipelines. For this reason, prefer either working with models that employ some corrective measures pertaining to the

required type of fairness. Also perform bias detection, and any necessary mitigation, on pretrained model outcomes.

Fair maintenance

After producing AI outputs, continue interacting with stakeholders to assert that their idea of fairness is correctly implemented. During cross-examination, keep a balance between justifying outputs as part of a fair process and accommodating constructive criticism. Do not over-rely on technical justification, as this would be another exhibition of principlism.

Throughout, try to make created systems explainable and follow processes that are understandable by the stakeholders; these are prerogatives of good scientific practices explored in the Which section. Furthermore, include the possibility for appeals against automated decisions, and be open to incorporating this feedback in maintenance, especially for longer-life systems. This can be part of the *TrustAIOps* employed for the fair design of new system versions.

Furthermore, implement processes that monitor the outputs of deployed systems and check if they work properly or reproduce biases. Most fairness formalism can be monitored with warning mechanisms akin to those of test-driven or metric-driven development. For example, check for distribution shifts in input data (which mirror changes in society) that either require greater corrective actions against already addressed biases or produce new forms of bias. Monitoring may involve continuous integration that asserts measure values both periodically and before deploying new system versions.

On the other hand, it can be hard to automatically check for adherence to non-mathematical fairness definitions. Aside from imposing certain processes by design, such as restricting access to specific system states or outputs, actively rechecking for non-technical definitions of fairness can be a lengthy process that slows down prototyping and

development, especially if non-technical people need to be consulted. As a middle ground, consider periodic personnel training and evaluation of your methodology.

What

is considered a fair outcome?

Goal
Learn about popular philosophical and mathematical fairness formalism.

Summary
Fairness is expressed either in philosophical terms, or as mathematical measures and constraints. The two fields adopt different outlooks but often arrive at similar principles. Assess your systems under many existing definitions and use the outcomes as examples to jump-start multistakeholder discussions.

Our suggested workflow gathers pragmatic stakeholder feedback, which you will use to model what is considered a fair outcome. To this end, work closely with social scientists to accurately transcribe gathered feedback to concrete fairness definitions. Pre-existing or new definitions that apply on AI systems are either philosophical and mathematical formalism explored in this section[6], or broader scientific standards and regulations that you will find in the Which section.

We already mentioned that you should enlist a wide range of fairness definitions to assess your systems at first and provide the assessment's outcomes to multistakeholder discussions (e.g., via presentations). For example, previous fairness measures can serve as examples of how to create new definitions by interlaying simpler concepts (Roy et al., 2023). In this process, not every type of prospective bias will be captured by your analysis, and not all found imbalances will reflect actual concerns. We reiterate that you should avoid principlism.

[6] Find more fairness formalisms in the Data Ethics Decision Aid (DEDA) handbook.

Do not select from the definitions of this section (including the philosophical ones) without listening to stakeholders.

Example
Consider our recurring mention to comparing false positive rates between groups, which is a type of disparate mistreatment described in the mathematical definitions below. This may create a false sense of fairness when the ground truth data used in the evaluation are also biased.

Here, we present fairness formalisms of two kinds: philosophical statements, and mathematical expressions. The former transcribe abstract concerns into actionable terms and are mainly used by social scientists, whereas mathematical expressions reflect the outcome of studying data or prediction imbalances from an algorithmic standpoint.[7]

Attempts to bridge the conceptual gap between the two fields map some philosophical terms to mathematical formulas and conversely. Table 1 provides such a mapping between the different kinds of definitions we present. Beware that interpreting philosophical definitions is controversial, which reinforces the need for stakeholder feedback. For example, the "equality of what?" debate (Matravers, 2002) created the schism between strict and luck egalitarianism.

Table 1. *Links between philosophical and mathematical definitions of fairness.*

Philosophical definition	Conceptually similar mathematization
Rawl's difference principle	Maximin; this is a direct implementation.
Strict egalitarianism	Do not use AI for basic rights, individual fairness.
Luck egalitarianism	Disparate mistreatment mitigation.
Deontic justice	Causal modeling, counterfactual fairness.
Representational fairness	Disparate impact.

[7] Bias in computer science is a broader term that also refers to deviations from expected values, and not only to fairness-related concerns.

Philosophical definitions

A philosophical framework for expressing fairness that is popular in social sciences is distributive justice. This examines how benefits and burdens should be distributed between members of society. For example, it could stipulate how decision-making or recommender AI systems should allocate goods, like food donations, burdens, like interest rates, or services, like access to healthcare. While following such definitions you can pursue either weak or strong fairness.

As a starting point, we provide three types of distributive justice to follow or adjust (Khan et al., 2021): luck egalitarianism, Rawl's difference principle, and strict egalitarianism. There also exist approaches to fairness that employ moral justification but do not fall under distributive justice. Two of them are deontic justice and representational fairness.

Types of distributive justice.

Luck egalitarianism (Arneson, 1990; Cohen, 1997) considers every person's starting point in society as the result of both the social situation into which they were born, and their biological potential. It thus pursues redistribution of burdens, goods, and services only where inequalities can be attributed to someone's starting point, and leaves in place inequalities which are the consequence of personal choices. In practice this could be expressed by bounding distributive outcomes to an acceptable range (Anderson, 2007). The algorithmic definition of explainable fairness (Shulner-Tal, 2022) also prevents the mitigation of biases that are the outcome of conscious decisions.

Rawls's difference principle (Rawls, 1971) justifies inequalities among members of society if they bring benefits even for the least advantaged. When following this principle by improving AI systems for everyone, take care to not disadvantage the least advantaged (Franke, 2021).

Mathematically
In practice, Rawl's difference principle translates to game theory maximin goals that maximize the worst system utility u_i across all individuals i (Ashrafian, 2023):

$$\text{maximize } \min_i u_i$$

System utility could correspond to the correctness of AI predictions, or to the latter's differentiable relaxation into a machine learning loss function (Rahmattalabi et al., 2019; Kang et al., 2022).

Strict egalitarianism (Carens, 1981) stipulates that every person has access to the same level of goods, burdens, and services. This may be at odds with making different predictions for each individual and often becomes the justification behind preventing AI from determining who gets access to basic human rights (Walzer, 2008), like the right to vote.

As an alternative to setting the predictions themselves as the subject of egalitarianism, some AI systems could be considered strictly egalitarian if they equalize or satisfy some derived property across all people. For example, procuring numeric errors between predicted and true scores across all individuals could be a type of burden that AI creators could try to equalize between each human data sample.

Mathematically
For utilities u_i of individuals i, strict egalitarianism may translate to zero variance between utilities:

$$\text{var}_i\{u_i\} = 0$$

Other moral justification.

Deontic justice (Parfit, 1997) is not concerned with an unequal situation per se, but rather with the way in which that state was produced, based on the combination of the historical and social context. Under this perspective, it is important to study the context itself and understand which groups of people have been discriminated against in the past and how forms of historical discrimination can affect AI systems in terms of fairness.

This type of justice does not lead to closed-form mathematical definitions of fairness but suggests that part of selecting them should involve consideration of the historic and contemporary factors responsible for broader systemic imbalances (Binns, 2018). Deontic analysis can also derive mechanisms that model the generation of bias; these mechanisms could be causal models (Pearl et al., 2016), for example used in the counterfactual fairness definition later.

Representational fairness (Binns, 2018) aims to fairly represent different identities, cultures, ethnicities, languages, or other social groups. In practice, it could be defined to achieve the same rate (or some other metric) of representation as the real-world groups. When uniform random sampling is performed on populations to gather datasets, this process becomes identical to disparate impact presented later in that it aims to create equilibria between representation rates over gathered data. However, dataset generation is rarely devoid of bias, at worst due to statistical noise, and the true (in the entire population instead of dataset) rates of representation need to be considered; dataset-derived rates may only serve as proxies when there is no other alternative.

When strong fairness is required, replicating the representation of oppressed groups may not be enough; fair representation may include

additional boosts for certain groups. Look at the example of boosting under-represented genders in hiring automation when strong fairness was first introduced.

Mathematical definitions

Mathematical fairness measures and constraints can be set as objectives of AI systems. Several such definitions exist in the literature (Castelnovo, 2021), with the most popular types being presented here; the list is non-exhaustive. Some of them are derived from early legal efforts at codifying fairness (Barocas and Selbst, 2016), but have since then been established as technical principles. For example, disparate treatment and disparate impact definitions originate from the namesake legal terminology of US Supreme Court decisions.[8]

Presented definitions offer a flexibility in analyzing slightly different contexts, which has led to the creation of comprehensive roadmaps on which to apply in certain kinds of known contexts (Ruf and Detyniecki, 2021). However, and contrary to their widespread use in research papers, they cover only a subset of known fairness concerns, much less the concerns that will arise in practice. As such, these definitions mostly serve as a basis from which to derive new ones.

Explored definitions are designed for weak fairness by imposing parity between individuals or groups of people under system performance measures of choice, respectively known as individual and group fairness. Still, their formulas are flexible enough to support strong fairness by

[8] *See:* Case Griggs v. Duke Power Co., 401 U.S. 424, 91 S. Ct. 849, 28 L. Ed. 2d 158 (1971) first recognized "disparate impact": https://www.publicjustice.net/what-we-do/access-to-justice/disparate-impact.
See for the conflicts between "disparate impact" and "disparate treatment": Ricci v. DeStefano, 557 U.S. 557, 129 S. Ct. 2658, 174 L. Ed. 2d 490 (2009).

placing different weights in the balance between groups. For example, given weak fairness that equalizes the fraction of positive predictions between population subgroups, strong fairness could boost the fraction for oppressed groups.

The greatest advantage of mathematical definitions of fairness is that they fit in existing algorithms, such as in the regularization of neural network backpropagation, and in the predictive constraints of linear programming. A different usage pattern is to create proxy definitions that apply on some components of an AI system pipeline (e.g., interventions to apply during preprocessing) that end up creating improvements under different – and potentially a wide range of – definitions at system outcomes. These improvements may be mathematically corroborated but measured only empirically.

In line with the potential incompatibility of different social contexts, not all mathematical definitions of fairness can be simultaneously satisfied. Various impossibility theorems (Kleinberg, 2016; Miconi, 2017) codify this statement for predictors that are neither infallible nor constant. To get an idea, the impossibility theorem provided by Miconi (2017) shows that non-perfect non-constant predictors cannot simultaneously satisfy the mathematical definitions of disparate impact, equalized odds, and predictive parity. Intuitively, reuse of the same base quantities in different bias assessment measures makes the latter optimized at different points.

Counterfactual fairness (Kusner et al., 2017) creates hypothetical scenarios where sensitive attributes assume different values while keeping all other attributes equal and stipulates that predictions should remain the same. Typical variations of counterfactual fairness consider mostly weak fairness that mitigates the effect of biases on AI predictions. But its underlying causal mechanisms can account for strong fairness assumptions by including the end-effect of AI on society

in their modeling. Usually, there is no process to verify the causal assumptions of the modeled bias generation, but such mechanisms can be agreed upon during stakeholder discussions; we already mentioned that extracting them is a type of deontic analysis.

Mathematically
For groups G, G' corresponding to different sensitive attribute values ($x \in G$ means that individual x assumes the corresponding sensitive attribute value) and predictions $\hat{y}_{x \leftarrow x \in g}$ that change which group $g \in \{G, G'\}$ individuals x belong to, counterfactual fairness is expressed in terms of probabilities $P(\cdot)$ as:

$$P(\hat{y}_{x \leftarrow x \in G} = y | x, x \in G) = P\big(\hat{y}_{x \leftarrow x \in g} = y | x, x \in G\big) \text{ for } g \in \{G, G'\}$$

Disparate impact (Kamiran and Calders, 2012) refers to disproportional representation within positive predictions.

Mathematically
For binary predictions $\hat{y} \in \{0,1\}$ of individuals x of groups G, G' corresponding to different sensitive attribute values, disparate impact is mitigated when*:*

$$P(\hat{y} = 1 | x \in G) = P(\hat{y} = 1 | x \in G')$$

The constraint of fully mitigating disparate impact is known as *demographic parity*, but rarely imposed as-is; mathematical relaxations create fractional comparisons or differences between the fractions, like the ⅘ rule, to quantify disparate impact mitigation and create objectives that can be optimized via machine learning. The above formula can also model other predictive tasks by reframing them as classification. For example, disparate impact becomes *proportional representation* if the predicted value for each data sample classifiers whether the sample resides within top-k recommendations (Zehlike, 2021).

Disparate impact tends to be at odds with improving AI accuracy, and on certain datasets its imposition can only make systems worse than producing a constant output (Pinzón, 2023). Recall that mathematical relaxations of disparate impact may be weaker than the namesake US legal term (Watkins et al., 2022).

Given that the causal modeling of counterfactual fairness involves latent variables that are statistically independent from the groups of people on which fairness is defined, there is an equivalence between counterfactual fairness over the entire population and disparate impact (Rosenblatt and Witter, 2023). However, even in this rare case where different definitions of fairness are compatible, they serve different purposes. In particular, causal mechanisms satisfy deontic explainability of biases, and can be used to make sense of the bigger picture, whereas disparate impact is easier to understand and work with as an objective.

Disparate mistreatment (Zafar et al., 2017) refers to dissimilar rates of making mistakes between protected and non-protected groups. For binary classifiers, where mistakes correspond to misclassification, disparate mistreatment often takes the form of equal false positive rates, false negative rates, precision, or recall.

Of these forms, false negative and true negative rate equality (which are equivalent to true positive and false positive rate equality) are also known as *equalized odds.*

Mathematically

For protected and non-protected groups G, G' for predictions $\hat{y}$ of data samples x with ground truth predictions y used for testing, this equalized odds satisfies both of these conditions:

$$P(\hat{y} = 1 | y = 1, x \in G) = P(\hat{y} = 1 | y = 1, x \in G')$$

$$P(\hat{y} = 1 | y = 0, x \in G) = P(\hat{y} = 1 | y = 0, x \in G')$$

The first of these conditions, which captures only false negative rate equality, is also known as *equal opportunity* (Hardt, 2016). Other misclassification measures may also be equalized or aggregated, for example across all decision thresholds of scoring systems (Gardner, 2019). All notions of disparate mistreatment rely on unbiased ground truth to compute errors, and we reiterate that this may be hard to find. Disparate mistreatment can be interpreted as a type of luck egalitarianism in that it attempts to equalize rates between different groups but does not concern itself with intra-group comparisons.

Disparate treatment (Peffer, 2009) refers to biases that arise from using sensitive attributes to make predictions, and originates from namesake US legislation. Removing sensitive attributes from the predictive process is also known as fairness through unawareness. This maps to the individualized symmetric treatment expression of weak fairness and is only applicable to predictions that utilize attribute values (e.g., attribute columns in tabular data, different classifiers per age group in images). However, it fails to account for social phenomena where attributes correlated with the sensitive ones. For example, if residence zip codes are strongly correlated to races, they can serve a proxy roughly representing the latter anyway.

Individual fairness (Dwork, 2012; Fleisher, 2021) constraints stipulate that similar individuals should be treated similarly.

Mathematically

Individual fairness constraints state that the distance d_Y between any two predictions y_1, y_2 should not exceed the distance d_X between corresponding individuals x_1, x_2. That is:

$$d_Y(y_1, y_2) \leq d_X(x_1, x_2)$$

This is also an example of individualized symmetric treatment. Defining how the distances should be computed (e.g., they can be proportional to p-norms) leads to different variations of individual fairness. Distance computations tend to create unintuitive interpretations for multidimensional data. This definition does not have a strong fairness counterpart, but is similar to certain expressions of strict egalitarianism in that it imposes a common property to be satisfied across all (pairs of) data samples.

Multi-fairness. Despite previously mentioned impossibility theorems, stakeholders tend to have multiple fairness concerns. Addressing these simultaneously is known as multi-fairness. This takes four main forms:

a) Satisfying multiple variations of the same type of fairness, as happens when disparate mistreatment is mitigated for both false positive rates and false negative rates.
b) Satisfying multiple types of fairness for the same protected attribute value corresponding to social groups or subgroups, as for example happens when aiming to protect women from disparate mistreatment while also adhering to individual fairness principles.
c) Analyzing sensitive attributes with multiple potential values, which correspond to multiple protected groups. These may compare each group with the rest of the population or compare all groups pairwise. For example, gender may not be a binary split between men and women but also include non-binary groups.
d) Considering the intersection of multiple sensitive attribute combinations (e.g., of gender and race) that correspond to several subgroups to be compared with each other.

To make matters more complicated, multi-fairness also covers combinations of the above, such as including different fairness measures for different attribute combinations. Two challenges of addressing multi-fairness are the aforementioned barrier between

conflicting definitions, and that combinations of multiple sensitive attribute values often partition data into many groups of few (e.g., zero) members that prevent statistically confident analysis. As of writing, addressing these challenges is still the subject of research.

Which
practices should be followed?

Goal
Learn about mandatory practices that should accompany fairness definitions.

Summary
Fairness frequently requires adherence to scientific rigorousness and human-centric design. Also follow international and local regulations. In many cases, regulations are at their infancy and may be updated in the future – likely following the EU's trustworthy AI framework.

In addition to philosophical and mathematical principles, definitions of fairness can include broader scientific and legal standards. Of these, scientific standards maintain rigorousness and ease of system use, and regulations consist of rules and recommendations laid out by the legal overseers of markets and services. Both kinds of practices are tied to the markets that AI systems affect and operate in. Thus, they may lie outside your control as a system creator.

Here, we provide a brief overview of practices that you may need to be aware of before entering new markets, such as scientific standards and regulation like the EU AI Act, so that you can work more easily within their boundaries once you start creating AI systems. The details provided are not applicable everywhere and may evolve over time. Much of the material is devoted to current EU regulation, as you will see that this has a worldwide effect. Work with legal experts to properly follow country-wide and international laws.

Scientific standards

Fairness may coalesce measures and constraints with standards of scientific quality, namely of rigorousness and of human-centric design. These are frequently recognized by regulations too. We present the latter separately because scientific standards lie closer to technical specifications in that, once selected, they are easy to follow by integrating state-of-the-art research outcomes in AI system creation.

Scientific rigorousness ensures that applied fairness definitions are achieved not only in exploratory scenarios under lab conditions, but also when AI is deployed in markets and starts affecting real people. Two cornerstones of rigorousness that are neglected in many systems, especially LLMs (Narayanan and Kapoor, 2023), are the ability to generalize and robustness. We now explore these concepts.

First, principles and system prediction correctness should generalize beyond training examples, that is, to new data. Balance between overfitting and underfitting with train-test-validation data splits and hyperparameter tuning. For example, neural networks with too few parameters will overgeneralize, but too many parameters will reduce testing correctness, especially for underrepresented groups. Critically examine default domain hyperparameters (including machine learning rates and the number, type, and breadth of neural network layers) before using them in new systems.

Additionally, AI systems should exhibit satisfactory correctness and adhere to fairness definitions for out-of-distribution data. Robust systems are those not easily misguided by small input perturbations or conscious (e.g., malicious) attempts at modifying predictions with input changes that are nigh-imperceptible to humans. In addition to safety against input perturbations and adversarial attacks, robustness should also safeguard against “poisoning” of training data to misguide AI

systems; this refers to attempts at instigating certain system behaviors through the data and could co-exist with real world social biases.

Human-centric design should accompany scientific progress to help both operators and stakeholders of AI systems. Fairness is already one aspect of such a design, but making systems friendly to humans is tied to more considerations. To begin with, data privacy is an important concern to garner trust from people whose data are being processed, especially throughout the EU, where privacy has been codified through the GDPR.[9] Data security directions can also include generative adversarial privacy that safeguards against reconstruction of user data; through this data security mechanism, the parameters of a generative model are extracted from the data themselves (Huang, 2017).

Furthermore, created systems should be easy to deploy and replicate so that their fairness can also be studied by independent third parties. To simplify deployment, follow good algorithmic design principles like benchmarking, replicability, and documentation. While creating AI, remain aware of the overarching fairness concerns arising from regulations and input from stakeholder.

Regulations

Integrating applicable AI legislation and ethical considerations in your systems is an essential process for market deployment. But the advantages of doing so also transcend to unregulated contexts, since fairness-oriented design and AI legislation are emerging worldwide (Jones et al., 2023). A market you should pay special attention to is the EU, where comprehensive and legal bidding regulation, especially the proposed AI Act[10], is expected to provoke global legislative efforts for AI

[9] *See:* General Data Protection Regulation (2016/679).

[10] In April 2021, the European Commission proposed the first EU regulatory framework for AI.

systems. This is called the *Brussels effect* and is expected to resemble the impact of GDPR, given that a substantial majority of providers put into service systems in the EU.[11] To help you acclimate to new markets that are influenced by this effect, you will later find an overview of legal tools for trustworthy AI in the EU.

Several countries introduced relevant legislation after the proposal of the EU AI Act in 2021 and can be considered part of the Brussels effect. Brazil proposed law *2338*[12] of May 2023 regarding AI. Canada proposed the *Artificial Intelligence and Data Act*[13] in June 2022. China issued the *Interim Measures for the Management of Generative AI Services* in July 2023 and the *Provisions on the Management of Algorithmic Recommendations in Internet Information Services* (China Law Translate, 2022). The US issued the *Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence* in October 2023.[14] AI fairness legislation was also adopted by some US states, like California, Connecticut, and Vermont (Engler, 2023; Cave and Paisner, 2023).

All the legal tools for AI systems worldwide are either in their initial stages, with their implementations yet unseen, or non-binding. Thus, let your collaborating legal experts continuously reassess deployed systems

[11] Article 2, European Commission, Proposal for a Regulation of The European Parliament and of the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts COM/2021/206 final.

[12] *See:* Bill nº 2.338/2023 available at: https://www25.senado.leg.br/web/atividade/materias/-/materia/157233.

[13] Part of the Digital Charter Implementation Act, 2022 (Bill C-27) available at: https://www.parl.ca/legisinfo/en/bill/44-1/c-27.

[14] Exec. Order No. 14110, 88 FR 75191, pp. 75191-75226 (October 30, 2023) available at: https://www.whitehouse.gov/briefing-room/statements-releases/2023/10/30/fact-sheet-president-biden-issues-executive-order-on-safe-secure-and-trustworthy-artificial-intelligence.

within regulatory fairness as the systems or laws change. This will often entail following a risk-based approach, where systems are re-examined on various aspects of trustworthiness upon changes. You can do this within the scope of fair design's *TrustAIOps*. Keep in mind that not all existing legal instruments that enforce rules aim to have the same effect, and each oversees different markets.

Respecting the legal provisions of regulations is mandatory, but implementing guidelines and recommendations is not. Nonetheless, satisfying a wide range of guidance for AI systems not only increases readiness for subsequent mandatory legal rules, but also contributes to business reputation and is considered an industry best practice. For example, NIST's *AI Risk Management Framework* (Tabassi, 2019) and its accompanying playbook rely on the US *National Artificial Intelligence Initiative Act*[15] but are intended to be voluntary, rights-preserving, non-sector-specific, and use-case agnostic for AI providers.

Another international non-binding legal attempt to support AI creators is the *Hiroshima AI process of the International Guiding Principles on Artificial Intelligence* (G7, 2023a) and the *Code of Conduct for Organizations Developing Advanced AI Systems* (G7, 2023b). This is based on the OECD *AI Principles*, and encourages entities worldwide to enforce detailed suggestions after identifying potential system risks.[16] Unlike the approach adopted by the EU AI Act, which classifies systems based on fairness-related risks, the Code of Conduct sets broader legal boundaries in relation to risks in certain critical sectors (e.g., nuclear risks, threats to democratic values, and human rights).

[15] US National Artificial Intelligence Initiative Act of 2020 (P.L. 116-283) available at: https://www.ai.gov/wp-content/uploads/2023/04/National-Artificial-Intelligence-Initiative-Act-of-2020.pdf

[16] *See*: no1 of the Hiroshima Process "International Code of Conduct for Organizations Developing Advanced AI Systems".

Trustworthy AI in the EU

In the EU environment, several measures apply when providers (notwithstanding their establishment) place on the market or put into service AI systems or their produced outputs.[17] The steps that you should follow as a system provider before deployment and operation can be broadly divided into two parts. First, the *Ethics Guidelines for Trustworthy AI* (AI HLEG, 2019) should guide all processes. Second, comply with the upcoming EU *AI Act* by identifying the level of risk of your systems. These two parts are explained in greater detail below.

Implement the internal protocols of the above two instruments to address existing or prospective issues regarding your systems' trustworthiness. Figure 5 presents effective tools that can help you. Of these, the *Assessment List for Trustworthy AI* (ALTAI) (European Commission, 2020) provides questions that – if correctly answered – let systems comply with the seven EU requirements for trustworthy systems. Then, *capAI* (Floridi et al., 2022) performs internal control for the *AI Act*. For the US, the playbook for the *AI Risk Management Framework* serves a similar purpose.

[17] *See:* Article 2, European Commission, Proposal for a Regulation of The European Parliament and of the Council laying down harmonized rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts COM/2021/206 final.

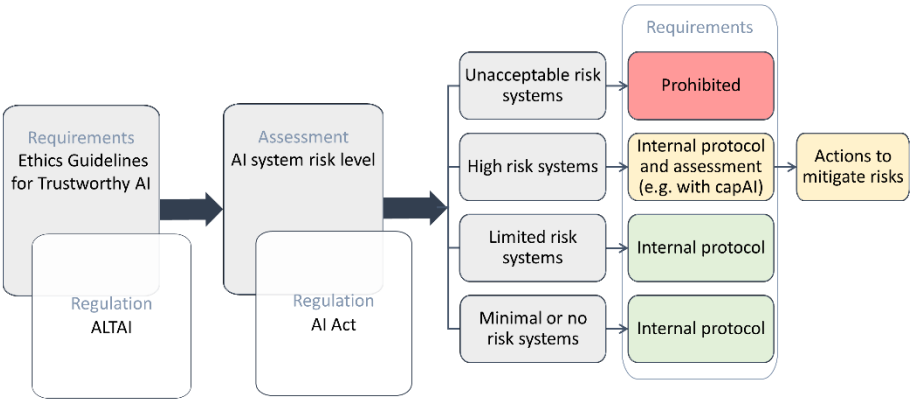


Figure 5. *The EU trustworthiness framework for AI system providers*

Ethics Guidelines for Trustworthy AI. The third chapter of these guidelines establishes an assessment list that comprises the following seven key requirements of trustworthy AI systems. You may recognize several of these from previously described scientific standards.

1. Human agency and oversight
2. Technical robustness and safety
3. Privacy and data governance
4. Transparency
5. Diversity, non-discrimination, and fairness
6. Societal and environmental well-being
7. Accountability

Ensuring ongoing system quality means continuous reassessment of these ethics guidelines. To this end, the *Independent High-Level Expert Group on Artificial Intelligence*, which was set up by the European Commission, has created the specialized *Assessment List for Trustworthy AI* (European Commission, 2020), which introduces proactive prevention of requirement violations.

EU AI Act. This is a framework for evaluating how much harm your systems might cause. Overall, it recognizes four categories of systems: a) *unacceptable risk systems* include real-time biometric systems (with a few exceptions for law enforcement purposes), social scoring algorithms, and AI that may involve manipulation risks, such as chatbots or deepfakes, b) *high-risk systems* are used for law enforcement, management of critical infrastructure, or recruitment, c) *limited risk systems* neither use personal data nor make any predictions influencing human beings, and include industrial applications in process control or predictive maintenance, *d) minimal or no risk systems,* such as AI-enabled video games or spam filters, are allowed free use. The European Commission expects most AI systems to fall into the category of limited risk systems (European Commission, 2023).

Given that unacceptable risk systems are outright banned, and that both limited and no risk systems do not require any action by providers, the weight falls on identifying and safeguarding high-risk systems. These require appropriate human oversight measures ("that cannot be overridden by the system itself and are responsive to the human operator, and that the natural persons to whom human oversight has been assigned have the necessary competence, training and authority to carry out that role") before being placed on the market or put into service[18].

The above-mentioned human oversight measures should be designed, assessed, and regularly reviewed by the system providers and, as a result, an assessment based on internal control would be efficient. In this direction, University of Oxford researchers provide the *capAI* tool (Floridi et al., 2022), which contains a procedure for assessing whether AI systems conform with the AI Act. Providers of high-risk systems can

[18] *See:* Recital 48 of the EU AI Act.

use this tool to demonstrate compliance while providers of low-risk systems can implement their commitment to voluntary codes of conduct. This tool consists of three components: a) an internal review protocol, which provides organizations with a tool for quality assurance and risk management, b) a summary datasheet to be submitted to the EU's future public database on high-risk AI systems in operation, and c) an external scorecard, which can also be made available to customers and other stakeholders (Floridi et al., 2022). Other tools can also assess and mitigate the risks of high-risk systems.

References

AI HLEG. “Ethics guidelines for trustworthy AI, Publications Office.” (2019)

Anderson, Elizabeth. "How Should Egalitarians Cope with Market Risks?." Theoretical Inquiries in Law (2007)

Arneson, Richard. “Liberalism, Distributive Subjectivism, and Equal Opportunity for Welfare.” Philosophy and Public Affairs, 19: 158-194 (1990)

Ashrafian, Hutan. "Engineering a social contract: Rawlsian distributive justice through algorithmic game theory and artificial intelligence." AI and Ethics 3, no. 4: 1447-1454 (2023)

Barocas, Solon, and Andrew D. Selbst. "Big data's disparate impact." California law review: 671-732 (2016)

Barocas, Solon, Moritz Hardt, and Arvind Narayanan. "Fairness in machine learning." Nips tutorial 1 (2017)

Beck, Kent, Beedle, Mike, van Bennekum, Arie, Cockburn, Alistair, Cunningham, Ward, Fowler, Martin, Grenning, James, Highsmith, Jim, Hunt, Andrew, Jeffries, Ron, Kern, Jon, Marick, Brian, Martin, Robert C., Mellor, Steve, Schwaber, Ken, Sutherland, Jeff and Thomas, Dave “Manifesto for Agile Software Development.” (2001)

Beauchamp, Tom L., and James F. Childress. Principles of biomedical ethics. Oxford University Press, USA (2001)

Biddle, Dan. Adverse impact and test validation: A practitioner's guide to valid and defensible employment testing. Routledge (2017)

Binns, Reuben. "Fairness in machine learning: Lessons from political philosophy." In Conference on fairness, accountability and transparency, pp. 149-159. PMLR (2018)

Calegari, Roberta, Gabriel G. Castañé, Michela Milano, and Barry O'Sullivan. "Assessing and enforcing fairness in the AI lifecycle." IJCAI International Joint Conference on Artificial Intelligence (2023)

Carens, Joseph. “Equality, Moral Incentives and the Market.” Chicago: Chicago University Press (1981)

Castelnovo, Alessandro, Riccardo Crupi, Greta Greco, Daniele Regoli, Ilaria Giuseppina Penco, and Andrea Claudio Cosentini. "The zoo of fairness metrics in machine learning." (2021)

Cavazos, Jacqueline G., P. Jonathon Phillips, Carlos D. Castillo, and Alice J. O’Toole. "Accuracy comparison across face recognition algorithms: Where are we on measuring race bias?." IEEE transactions on biometrics, behavior, and identity science 3, no. 1: 101-111 (2020)

Cave, Bryan and Paisner, Leighton. “2023 State-by-State Artificial Intelligence Legislation Snapshot.” Available at: https://www.bclplaw.com/en-US/events-insights-news/2023-state-by-state-artificial-intelligence-legislation-snapshot.html (2023)

China Law Translate. "Provisions on the Management of Algorithmic Recommendations in Internet Information Services." (2022)

Cohen, Gerald. “Where the Action is: On the Site of Distributive Justice.” Philosophy and Public Affairs, 26(1): 3-30 (1997)

Costanza-Chock, Sasha. “Design Justice. Community-led practices to build the worlds we need”, Cambridge, MA: The MIT Press (2020)

Dwork, Cynthia, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. "Fairness through awareness." In Proceedings of the 3rd innovations in theoretical computer science conference, pp. 214-226. (2012)

Engler, Alex. "The EU and US diverge on AI regulation: A transatlantic comparison and steps to alignment." Available at: https://www.brookings.edu/articles/the-eu-and-us-diverge-on-ai-regulation-a-transatlantic-comparison-and-steps-to-alignment (2023)

European Commission. "Regulatory framework proposal on artificial intelligence." Available at: https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai (2023)

European Commission, Directorate-General for Communications Networks, Content and Technology, "The Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self assessment." Publications Office. Available at: https://data.europa.eu/doi/10.2759/002360 (2020)

Fleisher, Will. "What's fair about individual fairness?." Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. (2021)

Floridi, Luciano, Matthias Holweg, Mariarosaria Taddeo, Javier Amaya Silva, Jakob Mökander, and Yuni Wen. "CapAI-A procedure for conducting conformity assessment of AI systems in line with the EU artificial intelligence act." Available at SSRN 4064091 (2022)

Franke, Ulrik. "Rawls's original position and algorithmic fairness." Philosophy & Technology 34, no. 4: 1803-1817 (2021)

Gardner, Josh, Christopher Brooks, and Ryan Baker. "Evaluating the fairness of predictive student models through slicing analysis." In

Proceedings of the 9th international conference on learning analytics & knowledge, pp. 225-234. (2019)

G7. “Hiroshima Process International Guiding Principles for Organizations Developing Advanced AI system.” (2023a)

G7. “Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems.” (2023b)

Gerards, Janneke, Mirko Tobias Schäfer, Iris Muis, and Arthur Vankan. "Fundamental Rights and Algorithms Impact Assessment (FRAIA)." (2022)

Giovanola, Benedetta, and Simona Tiribelli. "Weapons of moral construction? On the value of fairness in algorithmic decision-making." Ethics and Information Technology 24, no. 1: 3 (2022)

Hardt, Moritz, Eric Price, and Nati Srebro. "Equality of opportunity in supervised learning." Advances in neural information processing systems 29 (2016)

Himmelreich, Johannes. "Against “democratizing AI”." AI & Society 38, no. 4: 1333-1346 (2023)

Huang, Chong, Peter Kairouz, Xiao Chen, Lalitha Sankar, and Ram Rajagopal. "Context-aware generative adversarial privacy." Entropy 19, no. 12: 656 (2017)

John-Mathews, Jean-Marie, Dominique Cardon, and Christine Balagué. “From Reality to World. A Critical Perspective on AI Fairness”. J Bus Ethics 178, 945–959 (2022)

Jones, Fazlioglu, and Chaudry. "Global AI Legislation Tracker: by IAPP Research and Insights." In IAPP. IAPP AI Governance Center. Available at: https://iapp.org/media/pdf/resource_center/global_ai_legislation_tracker.pdf (2023)

Kamiran, Faisal, and Toon Calders. "Data preprocessing techniques for classification without discrimination." Knowledge and information systems 33.1: 1-33 (2012)

Kang, Jian, Yan Zhu, Yinglong Xia, Jiebo Luo, and Hanghang Tong. "Rawlsgcn: Towards rawlsian difference principle on graph convolutional network." In Proceedings of the ACM Web Conference 2022, pp. 1214-1225 (2022)

Khan, Falaah Arif, Eleni Manis, and Julia Stoyanovich. "Translation tutorial: Fairness and friends." In Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (2021)

Kleinberg, Jon, Sendhil Mullainathan, and Manish Raghavan. "Inherent trade-offs in the fair determination of risk scores." arXiv preprint arXiv:1609.05807 (2016)

Kong, Youjin. "Are “intersectionally fair” ai algorithms really fair to women of color? a philosophical analysis." In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, pp. 485-494. (2022)

Kusner, Matt J., Joshua Loftus, Chris Russell, and Ricardo Silva. "Counterfactual fairness." Advances in neural information processing systems 30 (2017)

Li, Bo, Peng Qi, Bo Liu, Shuai Di, Jingen Liu, Jiquan Pei, Jinfeng Yi, and Bowen Zhou. "Trustworthy AI: From principles to practices." ACM Computing Surveys 55, no. 9: 1-46 (2023)

Matravers, Matt. "Responsibility, luck, and the ‘equality of what?’ debate." Political studies 50, no. 3: 558-572. (2002)

Miconi, Thomas. "The impossibility of fairness: a generalized impossibility result for decisions." arXiv preprint arXiv:1707.01195 (2017)

Narayanan, Arvind, Sayash Kapoor. “Evaluating LLMs is a minefield” Princeton University. Available at: https://www.cs.princeton.edu/~arvindn/talks/evaluating_llms_minefield (2023)

OECD. “Recommendation of the Council on Artificial Intelligence.”, OECD/LEGAL/0449 (2023)

O’Neil, Cathy, and Hanna Gunn. "Near-term artificial intelligence and the ethical matrix." Ethics of Artificial Intelligence: 235-69. (2020)

Parfit, Derek. “Equality and priority.” Ratio, 10(3): 202-221 (1997)

Pearl, Judea, Madelyn Glymour, and Nicholas P. Jewell. Causal inference in statistics: A primer. John Wiley & Sons, (2016)

Peffer, Shelly L. "Title VII and disparate-treatment discrimination versus disparate-impact discrimination: The Supreme Court’s decision in Ricci v. DeStefano." Review of Public Personnel Administration 29.4: 402-410 (2009)

Pinzón, Carlos, Catuscia Palamidessi, Pablo Piantanida, and Frank Valencia. "On the incompatibility of accuracy and equal opportunity." Machine Learning: 1-30 (2023)

Rahmattalabi, Aida, Phebe Vayanos, Anthony Fulginiti, Eric Rice, Bryan Wilder, Amulya Yadav, and Milind Tambe. "Exploring algorithmic fairness in robust graph covering problems." Advances in neural information processing systems 32 (2019)

Rana, Saadia Afzal, Zati Hakim Azizul, and Ali Afzal Awan. "A step toward building a unified framework for managing AI bias." PeerJ Computer Science 9: e1630 (2023)

Rawls, John. “A Theory of Justice”, Harvard, MA: Harvard University Press (1971)

Rosenblatt, Lucas, and R. Teal Witter. "Counterfactual fairness is basically demographic parity." In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, no. 12, pp. 14461-14469 (2023)

Roy, Arjun, Jan Horstmann, and Eirini Ntoutsi. "Multi-dimensional Discrimination in Law and Machine Learning-A Comparative Overview." In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, pp. 89-100 (2023)

Ruf, Boris, and Marcin Detyniecki. "Towards the right kind of fairness in AI." arXiv preprint arXiv:2102.08453 (2021)

Ruggieri, Salvatore, Jose M. Alvarez, Andrea Pugnana, and Franco Turini. "Can we trust fair-AI?." In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, no. 13, pp. 15421-15430 (2023)

Saxena, Nripsuta Ani, Karen Huang, Evan DeFilippis, Goran Radanovic, David C. Parkes, and Yang Liu. "How do fairness definitions fare? Examining public attitudes towards algorithmic definitions of fairness." In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, pp. 99-106 (2019)

Shulner-Tal, Avital, Tsvi Kuflik, and Doron Kliger. "Fairness, explainability and in-between: understanding the impact of different explanation methods on non-expert users’ perceptions of fairness toward an algorithmic system." Ethics and Information Technology 24, no. 1: 2 (2022)

Tabassi, Elham. "Artificial Intelligence Risk Management Framework (AI RMF 1.0)." NIST Trustworthy and Responsible AI, National Institute

of Standards and Technology, Gaithersburg, MD, https://doi.org/10.6028/NIST.AI.100-1 (2023)

Thakur, Vishesh. "Unveiling gender bias in terms of profession across LLMs: Analyzing and addressing sociological implications." arXiv preprint arXiv:2307.09162 (2023)

Whittaker, Meredith, Meryl Alper, Cynthia L. Bennett, Sara Hendren, Liz Kaziunas, Mara Mills, Meredith Ringel Morris et al. "Disability, bias, and AI." AI Now Institute 8 (2019)

Xiang, Alice, and Inioluwa Deborah Raji. "On the legal compatibility of fairness definitions." arXiv preprint arXiv:1912.00761 (2019)

Watkins, Elizabeth Anne, Michael McKenna, and Jiahao Chen. "The four-fifths rule is not disparate impact: a woeful tale of epistemic trespassing in algorithmic fairness." arXiv preprint arXiv:2202.09519 (2022)

Walzer, Michael. Spheres of justice: A defense of pluralism and equality. Basic books (2008)

Wachter, Sandra, Brent Mittelstadt, and Chris Russell. "Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI." Computer Law & Security Review 41: 105567 (2021)

Weinberg, Lindsay. "Rethinking fairness: an interdisciplinary survey of critiques of hegemonic ML fairness approaches." Journal of Artificial Intelligence Research 74: 75-109 (2022)

Zafar, Muhammad Bilal, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. "Fairness beyond disparate treatment & disparate impact: Learning classification without disparate

mistreatment." In Proceedings of the 26th international conference on world wide web, pp. 1171-1180 (2017)

Zehlike, Meike, Ke Yang, and Julia Stoyanovich. "Fairness in ranking: A survey." arXiv preprint arXiv:2103.14000 (2021)

About the Authors

Emmanouil Krasanakis* is a PostDoc research associate at CERTH-ITI. He holds a PhD degree in Electrical & Computer Engineering from the Aristotle University of Thessaloniki, and a BSc degree from the same faculty. He has been involved in nine EU-funded projects and is the co-author of seven journal articles and eleven conference papers. His main research interests lie in graph theory and graph neural networks, machine learning with focus on algorithmic fairness and discrimination, and software engineering.

Marta Gibin is a research fellow at the University of Bologna. She holds a PhD in Sociology and Social Research from the same university and has been involved in the EU-funded projects MAMMOth and ONCORELIEF. In 2022, she won a research grant from the European Society for Health and Medical Sociology for a visiting period at the Datafied Life Collaboratory of the University of Helsinki. Her research interests mainly concern the social implications of the use of AI systems, with a focus on their implementation in the healthcare sector.

Stavroula Rizou is a PostDoc research associate at CERTH-ITI. She holds a PhD from the University of Macedonia on "Cross-border transfer of personal financial data: legal approach", for which she was honoured with a grant from the Hellenic Foundation for Research and Innovation. She also holds a research award from the Special Account for Research Funds of the University of Macedonia. She graduated from the Faculty of Law of the Aristotle University of Thessaloniki and holds a MSc. Her research interests focus on the interaction of data protection with innovative IT applications and artificial intelligence legislation.

* *Corresponding author. Email:* maniospas@iti.gr.

MAMMOth consortium members collaborated on writing this guide by reviewing it and discussing finer points with the authors. Involved members and their affiliations are listed below in alphabetical order:

Ana Maria Jaramillo	CSH
Arjun Roy	UniBw
Christos Koutlis	CERTH-ITI
Dagmar Heeg	University of Groningen
Elli Nikolakopoulou	IASIS NGO
Evren Yalaz	Trilateral Research
Fariba Karimi	CSH
Symeon Papadopoulos	CERTH-ITI
Ian Slesinger	Trilateral Research
Ilias Michail Rafail	IASIS NGO
Ioannis Sarridis	CERTH-ITI
Mauritz Cartier van Dissel	CSH
Nathan Ramoly	IDnow
Thanos Loules	IASIS NGO
Swati Swati	UniBw

External reviewers provided additional feedback that let us refine this guide. Reviewers and their affiliations are listed below in alphabetical order:

Konstantinos Zacharis	UniBw
Tai Le Quy	l3s
Tobias Callies	UniBw
Siamak Ghodsi	l3s