

AI-fairness: the FAIRBRIDGE approach to practically bridge the gap between socio-legal and technical perspectives

Giovanni Ciatto*, Mattia Matteini*, Laura Sartori†, Maria Rebrean°, Catelijne Muller°, Andrea Borghesi*, and Roberta Calegari*

*Dept. of Computer Science and Engineering (DISI), ALMA MATER STUDIORUM—Università di Bologna, Italy
{giovanni.ciatto, mattia.matteini, andrea.borghesi3, roberta.calegari}@unibo.it

†Dept. of Political and Social Sciences, ALMA MATER STUDIORUM—Università di Bologna, Italy
l.sartori@unibo.it

°ALLAI, Netherlands
{maria.rebrean@allai.nl, catelijne.muller}@allai.nl

Abstract

Addressing the need for AI systems free from discrimination requires a multidisciplinary approach that combines social, legal, and technical perspectives. Despite significant advancements in research and technical solutions, a gap remains between socio-legal and technical approaches. This paper proposes a meta-methodology – namely, FAIRBRIDGE – to bridge this gap, offering a reference for defining AI fairness methodologies that integrate all three perspectives. The meta-methodology utilizes a questionnaire-based system where socio-legal and technical domain experts iteratively refine questions and responses, supported by automation.

Keywords: FAIRBRIDGE, AI-fairness, methodology, practice, Q/A

1. Introduction

Modern computational systems are becoming increasingly complex and pervasive, mostly due to the ever-increasing capabilities of Artificial Intelligence (AI) technologies. As AI grows in autonomy and performance, it also starts affecting human lives, in domains such as healthcare, justice, education, finance, etc. Unfortunately, applications of AI in human-centered domains are revealing how AI-powered systems tend to reproduce, and sometimes amplify, the biases present in the data they are trained on, or in the people who design them.

By AI system, we mean any Software (SW) system involving AI approaches, either those that require some data-driven training phase (e.g., Machine Learning (ML), models) or methods from the learning-free AI area (e.g., agent-based systems, optimization, logic

programming, etc). Such systems may be affected by biases or discrimination, either due to the content of the data or in the way it was collected, as well as in the way the algorithm was designed or developed.

To mitigate this issue, recent efforts in AI-fairness research have been focusing on either (i) developing statistical algorithms for detecting and mitigating biases, or (ii) defining guidelines and best practices for ensuring fairness in AI systems. In the former, researchers provide practical methods for addressing biases either in the training data or in the training process. Each method is usually tailored to a specific type of bias, dataset, AI algorithm/task, etc., and practitioners are left to navigate a vast and fragmented literature to find the right tool for their specific case. Conversely, in the latter case, researchers provide general guidelines and lists of best (or worst) practices – in the form of legal articles, ethical principles, social norms, desirable outcomes, etc. –, but these are often too abstract, and practitioners are left to interpret and apply them to their specific cases.

Summarising, further research is needed to amalgamate statistical methods (i.e., the technical instruments for fixing AI) and guidelines (i.e., the criteria about if, when, and how those tools should be used), and any effort in this direction should involve a multidisciplinary team of experts, including sociologists, legal experts, statisticians, and computer scientists. Along this line, our goal is defining Fair-by-Design (FbD) *methodology* (Manjarrés et al., 2021), which encompasses all the aforementioned fields of expertise (Shneiderman, 2020). Here, FbD refers to an approach to the construction of AI system where fairness considerations are integrated into the AI lifecycle from the outset. We first attempted to let experts from different domains converge on a shared methodology—soon realising how this approach is

unscalable due to the vastness of the literature and the time required to align all experts' knowledge. Consequently, we designed an incremental approach, which is the main contribution of this work.

We propose a *meta-methodology* for fairness engineering, consisting of a *stable* set of core principles and an *evolvable* pool of practices for steering end users towards a deeper understanding of the problem/domain they are dealing with, and for guiding their decision-making. The meta-methodology should then be reified into a guidelines-provisioning SW system – namely, FAIRBRIDGE – whose capabilities and degree of automation can be incrementally improved, as prescribed by the meta-methodology itself.

FAIRBRIDGE should not replace human decision-makers, but rather assist them in making informed, conscious, and timely decisions about the fairness of their AI systems. Moreover, we also aim to involve users in the process as much as possible, to make them aware of the decisions they are making—as opposed to forcing them to take a predetermined path. To make this possible, we constrain FAIRBRIDGE to be based on a Question–Answering (Q/A) mechanism, where questions are meant to help users in making fair AI systems, step by step, and answers are collected to improve the system's recommendations, dynamically. FAIRBRIDGE allows for answers to be recommended by the system itself, but only when recommendations are based on clear and sound practices, that a human expert may validate.

Despite its simplicity, this approach is highly preferable because of its incremental nature. The exact amount, phrasing, and order of the questions, as well as their admissible answers, can be incrementally refined by experts in law, sociology, statistics, and computer science as part of their research activities. Recommendations may be initially delegated to human experts, but they may also be automated (or semi-automated) later on. Accordingly, the remainder of this paper is structured as follows. In Section 2, we provide some background on the state of the art in AI-fairness research, and we summarise the most relevant related works in the field. In Section 3, we introduce our meta-methodology for fairness engineering, formalising requirements, actors, and design principles. Next, in Section 5 we demonstrate the instantiation of the meta-methodology.

2. Background

The integration of social, legal, ethical, and technological perspectives comes with two challenges: complexity and interdisciplinarity. Each perspective

operates within its framework. Social, legal, and ethical perspectives focus on human behaviour, ethical principles are designed for digitalization and regulation, while technological perspectives prioritise efficiency, functionality, and innovation. While bridging these perspectives requires interdisciplinary collaboration (Shneiderman, 2020), finding approaches with this aim is tough. To the best of our knowledge, our effort is the first of its kind.

2.1. AI Lifecycle

Considering the entire AI lifecycle is crucial for assessing fairness as biases may infiltrate at different stages (Caton & Haas, 2024). Analogies exist among the socio-legal and technical perspectives: for instance, all of them identify the necessity of (i) carefully considering how data has been sampled (collected or generated), (ii) analyzing whether it contains any bias, and (iii) processing it without introducing further bias. They also emphasise how AI systems should be evaluated for fairness.

Differences exist as well. On the one side, the technical perspective focuses on a narrower AI lifecycle compared to the social-legal one. In other words, fewer phases are considered by technological approaches (Calegari et al., 2023). The technical perspective offers a detailed approach to the “development” and “evaluation” phases in the socio-legal lifecycle, examining intervention levels for making AI systems fair. It emphasizes pre-processing, in-processing, and post-processing phases. Very few technical works try to broaden the perspective. for instance,

On the other side, social and legal approaches focus on “building blocks” for fair AI, such as risk assessment, stakeholder identification, regulatory analysis, and fundamental human rights impact assessment—the “Scoping” and “Risk Analysis” phases of the AI lifecycle. These building blocks provide insights into developing and using AI systems fairly, identifying gaps, and ensuring equitable outcomes. These aspects are partly covered in the “Planning” stage of the technological lifecycle (“Business requirements” and “Analytics approaches”).

Concerning fundamental rights impact assessments, these will be legally required for some AI systems, yet no standard for implementing them has emerged so far. Generally speaking, it must be noted that the interplay between sociological/legal and technological perspectives is still in its infancy: engineering solutions tend to adopt excessively reductionistic approaches (discarding the big picture of social, economic, and institutional constraints). In contrast, various

sociological/legal indications and suggestions struggle to coalesce into well-defined and actionable guidelines that can be applied in practice (Wachter et al., 2021).

2.2. Practical Issues

Some phases of the AI lifecycle may only be covered by guidelines, descriptive methodologies, or standards that explicitly outline how to address fairness compliance from a social-legal perspective. The more technical phases involving the selection of a “fair” algorithm for the specific case and the “best” fairness metric for the particular case need to be “guided” by the social and legal aspects (de Almeida et al., 2021). Socio-legal approaches offer broad guidelines without defining practical fairness measurements, leaving interpretation to technical experts or courts. In contrast, technical approaches aim to define fairness metrics (Mitchell et al., 2021). Two elements are essential for fairness awareness: (i) defining fairness notions from social, legal, ethical, and technical perspectives, and (ii) providing a quantitative mechanism to measure them where possible.

Fairness notions vary by *context* and stakeholders, requiring different activities for fulfilment. They can be measured using statistical formulas (fairness metrics), but this leads to many metrics, each measuring slightly different aspects of fairness. Recent efforts, like the ontology of (Franklin et al., 2022), aim to better organise these metrics by defining them, describing their use cases, and detailing their relationships. Legal and social perspectives on fairness are case-dependent, interpreted differently, and influenced by social and institutional factors; setting thresholds for what is fair or unfair. Current FbD practices have a clear gap due to the challenges of complexity and *interdisciplinarity* in integrating multiple perspectives. Bridging these perspectives requires interdisciplinary collaboration.

These insights can be summarized as follows: (i) FbD approaches require the collaboration of an interdisciplinary team; (ii) a FbD methodology should be tailored according to the context. For this purpose, we propose a meta-methodology rather than a single methodology; we want to provide tools for *building fair methodologies*. Both socio-legal and technical experts shall operate this tool. Our approach is described in the following section.

3. Fair-by-Design (FbD) Meta-Methodology

The goal of FbD methodologies is to ensure that AI systems are fairly designed and developed, addressing potential biases or discrimination since the

very beginning of the design and development phases, so that the final system is inherently fair—other than, of course, effective and efficient in whatever AI task it is designed for.

Turning this vision into reality is not straightforward, for several reasons. First, the very notion of ‘fairness’ depends on the socio-technical context in which the AI system is (going to be) deployed, as proved by the many fairness notions proposed in the literature. In fact, there exist many different notions and definitions of fairness (Mehrabi et al., 2021), ranging from distributive to procedural, and encompassing group, individual, and causal fairness. We must also note that different notions can be more well-suited to different use cases (Makhlouf et al., 2021) – it is impossible to find a fairness notion that can be applied in all circumstances. Hence, we do not target a single definition of fairness, but we rather let the Q/A mechanism guide the selection of the most suitable one (or multiple ones) according to the specifics of the use case onto which the meta-methodology has to be tailored.

To complicate the matter, contexts are not set in stone, and they are most likely going to evolve. Second, the technicalities involved in the development of AI systems are complex, and inherently domain-specific: when fairness is added to the mix, the complexity increases even further, as AI technicians need to translate social, legal, and ethical requirements into technical requirements—and the process can hardly be automated. Third, when brought to practice, the AI workflow is non-linear and iterative, and the situation where an AI workflow is started completely from scratch is rare: most commonly, AI systems are developed out of existing datasets and algorithms—with fairness considerations being added as an afterthought.

For all these reasons, *practical* attempt to build an FbD methodology should keep into account the following desiderata:

(a) the methodology should consider the cultural context and the domain in which the AI system is going to be applied; (b) the methodology should adapt to any change in the cultural context as it evolves; (c) the methodology should assist experts in the activity of translating the social, legal, and ethical requirements into technical requirements but (d) without replacing the human decision-maker; and, finally, (e) the methodology should account for pre-existing datasets and algorithms as the basis for the fair AI system to be developed.

To effectively address the aforementioned desiderata, we propose following an incremental approach, where the FbD methodology is developed incrementally, starting from an initial version to

be repeatedly refined. Accordingly, we identify a stable set of core principles, which should *not* vary among the different versions of the methodology, and a pool of practices by which the novel versions of the methodology can be attained. In this way, we are essentially proposing a *meta-methodology* for fairness engineering, by which context-specific FbD methodologies can be developed and refined over time. Such a reiterated approach is paramount to ensure that the resulting methodologies are kept up-to-date as the context evolves—as prescribed by desideratum (b) above.

In practice, we want to provide a tool (the meta-methodology) to build and adjust FbD methodologies for AI. This degree of separation is important as while the fairness methodology depends on the context and needs to be tailored to its specific domain, the meta-methodology can instead be general and shared across different social and legal contexts.

Principle 1: Question–Answering. We require an FbD methodology to be based on a Q/A mechanism, where questions are meant to help decision-makers in either building fair AI systems from scratch or detecting and mitigating biases in partially-developed AI systems.

Principle 2: what Q/A and why? The questions and their admissible answers should be designed to deepen decision-makers understanding of the problem and the domain they are dealing with, and to make them aware of any relevant issues concerning their application scenario—hence guiding their decisions accordingly. For this reason, the questions and answers should be designed by experts in law, sociology, statistics, and computer science, as prescribed by desideratum (a) above.

Principle 3: order of questions. Furthermore, to comply with desideratum (c), questions should be hierarchically presented to decision-makers, where most general questions are asked first, and more specific questions are asked later. Also, the answers to earlier questions may impact which and how many questions are asked later on. Ideally, by the end of the questionnaire, each fairness-related aspect should have been considered and addressed by decision-makers.

Principle 4: decision support. To comply with desideratum (d), the system should avoid giving answers in place of decision-makers. Yet, to answer the questions and, consequently, make decisions, decision-makers may need additional information (e.g., best-practices, summaries of numerical data coming from the available datasets, charts or tables concerning algorithms under training, etc.). For this reason, the methodology should ease access to such information in a timely fashion. In particular, while filling out

the questionnaire, decision-makers should be able to request additional information from experts, and let them *suggest* an informed answer to the question at hand.

Principle 5: starting point. Finally, to account for desideratum (e), the methodology should involve *early* questions aimed at discriminating among one of three possible initial situations: (i) decision-makers have already collected some data and developed an AI system on top of that data, (ii) decision-makers have already collected some data but have not yet developed an AI system on top of that data, (iii) decision-makers have not yet collected any data nor developed any AI system.

3.1. The FAIRBRIDGE System

It is impossible to apply the same set of questions to all situations, as some might be relevant only in specific fields. For instance, depending on the application domain, the stakeholders being involved by the AI-system under development may vary significantly.

The selection of the stakeholders is a difficult step in itself, and its precise definition is outside of the scope of this paper. To identify the stakeholders, a series of questions must be answered *before* the application of the FAIRBRIDGE approach. The questions are divided into three categories to identify: (i) stakeholders affected by the AI-system (who benefits or is harmed by the unfairness?): the impact can be positive or negative and direct or indirect; (ii) stakeholders that have power over the development and deployment of the AI-system (who decides on AI use? who manages, governs, audits, regulates, or supervises the AI system?) (iii) stakeholders that have information that would aid with the development of a fair AI-system (who develops the AI system? who has expertise in the system’s domain? Who will use or interact with the AI system?).

So, our proposed solution is to develop a *software* system for fairness engineering, namely the FAIRBRIDGE system, which would make our FbD methodology viable in practice for decision-makers, while allowing for experts to iteratively expand and refine the pool of admissible questions and answers. Such a system is conceptualised in Figure 1.

As highlighted in the figure, FAIRBRIDGE assumes the involvement of several actors, and requires that experts in law, sociology, statistics, and computer science fill the system with questions and answers, so that decision-makers can be supported in making fair decisions about their AI systems.

3.1.1. The Actors We start by modeling the actors that affect and are affected by the system itself.

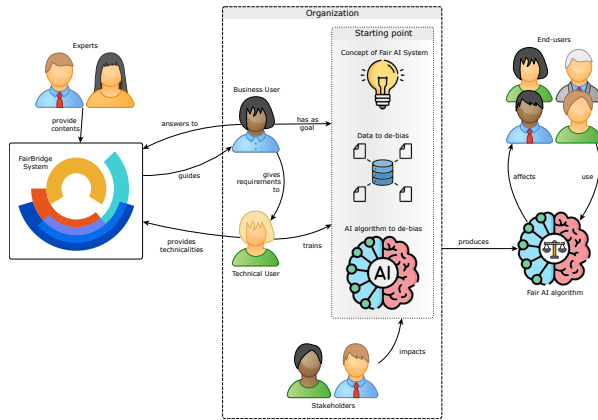


Figure 1. Concept of the FairBridge approach to fairness engineering.

FAIRBRIDGE is intended to be used by *organizations* whose goal is to develop fair AI algorithms¹, which are expected to be used by some *end-users*. Another perspective on the same plot is that end-users are potentially *affected* by the AI algorithms developed by the organizations, thus their AI algorithms need to be fair for those end-users. End users could be private individuals or companies as well as public institutions or civil society organizations willing to use the final AI system. In the eyes of FAIRBRIDGE, end-users may also include “affectees”, i.e., people affected by (subjected to) the decision of an AI system being used by somebody else—e.g., job applicants, or welfare recipients.

The actual users of FAIRBRIDGE are the *members* of the organizations who are responsible for developing the AI systems. These members are divided into two categories:

- **Business users** are responsible for any decision concerning the target AI system. As such, they are in the position of making decisions and should answer the questions posed by FAIRBRIDGE. For this reason, they should have sufficient background knowledge to understand the questions and the admissible answers, or know who to ask for help when this is not the case.
- **Technical users** are responsible for the implementation of the AI system, following the decisions taken by the business users. They are SW developers and data scientists, and they possess adequate technological expertise to develop AI systems. Technical users are responsible for providing technicalities and

¹Possibly composed into systems

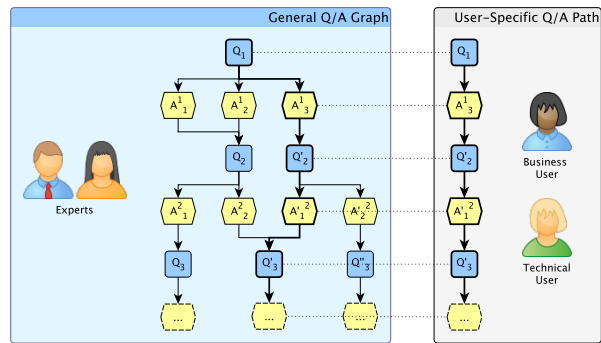


Figure 2. Graphical depiction of the Q/A mechanism, as it is perceived by experts (i.e. a graph) and by business/technical users (i.e. a path).

details to the FAIRBRIDGE system, to support business users in answering the questions.

Organizations may also include further actors that have an impact on the target AI system and its fairness. We refer to these actors as *stakeholders*².

The Q/A mechanism includes questions aimed at identifying the stakeholders, in such a way their existence and views can be included in any fairness-assessment and enforcing action. A similar argument holds for the *potential* end-users of the target AI system, whose profile should be identified by the Q/A mechanism as early as possible.

Finally, FAIRBRIDGE involves *experts*, i.e., people knowledgeable in law, sociology, statistics, or computer science, and therefore able to design the questions and answers to be included in the system. These experts could be policymakers, regulators, researchers, or even practitioners, as long as they have the required expertise.

Technical experts (the AI system developers) might have diverging goals from the end-users of the system. We explicitly involve social, ethical, and legal experts in our approach to act as a liaison with the end-users and to ensure the resolution of eventual conflicts, via a constant dialogue stimulated by the action of answering the questionnaire.

3.2. The Q/A Mechanism

The Q/A mechanism is the core of FAIRBRIDGE, and it is the tool by which the meta-methodology is reified into a practical SW system. It involves a set of *relevant* questions and their *admissible* answers, plus a partial ordering relation, which defines the order in which the questions should be asked to the business

²These could be managers, customers, suppliers, as well as any person or group of people affecting the development process with their biases or diverging interests.

users. The answer to a question may impact which and how many questions are asked later on to the same business user. This situation is depicted in Figure 2, where the Q/A mechanism is represented as a *graph* (for experts) and as a *path* (for business/technical users). Even if the graph is the same for all business users, each business user may follow a different path, depending on their particular use case, domain, goals, and constraints. In other words, the graph represents the whole set of questions and answers, and their ordering, while the path represents the subset of questions and answers that are shown to one particular business user.

The Q/A graph is designed, filled, and refined by experts. Conversely, Q/A paths are constructed by business users and technical users who are using FAIRBRIDGE, and they are tailored to the specific needs of the organization they belong to.

The Q/A Graph.

The Q/A is the knowledge base of FAIRBRIDGE. In abstract terms, it can be modelled as a Directed Acyclic Graph (DAG), where: (i) **nodes** represent either questions or answers; (ii) **questions** are free-text strings, containing a natural language sentence expressing an *inquiry*, plus an identifier (unique within the whole graph) such as Q1, Q2, etc. They also contain a type descriptor, which indicates how the question should be answered, e.g., whether it is a single- or multiple-choice question, a free-text question, etc. (iii) **Answers** are free-text strings, containing a natural language sentence expressing one possible *response* to some question, plus an identifier (unique w.r.t. the same question) such as A1, A2, etc. Combined with a question's identifier, answers' identifiers form a unique identifier of the answer within the graph (e.g., Q1-A1, Q1-A2, etc.). (iv) **Edges** are of two sorts: either *question-to-answer* edges, denoted by $Q \mapsto A$, or *answer-to-question* edges, denoted by $A \mapsto Q$. Edges of the former sort ($Q \mapsto A$) represent the fact that a question Q has an admissible answer A , whereas edges of the latter sort ($A \mapsto Q$) represent the fact that question Q should follow when answer A is selected. There cannot exist an edge mapping two questions or two answers. We assume that each question has *at least one* outgoing edge (to an answer), and each answer has *exactly one* incoming edge (from a question). (v) The **root** of the Q/A graph represents the first question to be asked to each business user.³ (vi) The **leaves** of the Q/A graph represent the closing answers to be asked of each business user, as the last ones. Technically speaking, a leaf is an answer node with no *outgoing* edges. We assume that the Q/A graph is *connected*: each questionnaire is guaranteed to have a

³The root is a question node with no *incoming* edges. We assume the root to be unique

beginning and an end, and each question is guaranteed to have (at least) one admissible answer.

When filling the Q/A graph, experts should take into account the socio-technical context in which the AI system is going to be deployed, and phrase the questions and answers so they are understandable to business users. They should also engineer the order in which the questions are asked, possibly starting from the most general ones, and proceeding to the most specific ones. Branches are possible and welcome, and this is key to the engineering of the Q/A graph in the long run.

About Q/A Paths. The Q/A graph shall eventually be presented to business users as a linear questionnaire. This implies that some root-to-leaf *path* of the graph will be navigated by a business user. In the eyes of the business user, the questionnaire is a sequence of questions and answers, that they construct *interactively* by selecting the admissible answers to the questions they are presented with. Interactivity is the key concept: the questions and admissible answers should be presented so that the business user is guided in taking the correct fairness-related decisions, at the right time.

To make this possible, FAIRBRIDGE should expose a (possibly graphical) interface to business users, essentially consisting of two major operations. First, **get the next question**, aimed at presenting the next question to the business user, depending on the previous answers. The first time this operation is triggered, the root question is presented to the business user. Afterwards, the operation should return the question following the last answer given by the business user in the Q/A graph. Second, **answer the question**, essentially aimed at letting the business user select one or more admissible answers to the question they are presented with or fill (in the free-text case) with custom text. Business users may require taking *informed* decisions, e.g. decisions based on statistical or technical operations which require ad-hoc expertise. To this end, FAIRBRIDGE allows technical users to support filling out the questionnaire.

On the role of technical users. When decisions to be taken by a business user involve some technical operations, such as computing statistical metrics, or training an AI model on some data, questions (and answers) should be designed to let the business user choose *which* technical operation to perform, and *when*, yet the actual operation should be performed by some technical user. The outcome of the technical operations should be presented to the business user, possibly as part of the questionnaire, so that they can make an informed decision. This may be the case for instance a question like: “*Will end-users be discriminated on a racial basis by the AI system under development?*” and admissible answers could be as simple as “*Yes*” or “*No*”: yet, if

the training data is already available, the technical user may exploit a fairness metric such as *disparate impact* to suggest which answer will be more likely to be correct. In this case, the exploitation of the disparate impact metric is a technical operation, involving the execution of some code on the available data.

To support this kind of contextual suggestion, FAIRBRIDGE should expose an ad-hoc Application Programming Interface (API) to technical users, allowing them to react to: (1) **question asked**, triggered when a question is presented to the business user for the first time; (2) **answer given**, triggered when the business user provides answers to some questions. In reaction to these events, SW scripts written by the technical users may perform any kind of computation (from data analysis to training AI models).

Technical users' scripts may also affect the questionnaires, via: (i) **Tag an answer** *A* for the current question as either recommended or discouraged, because of motivation *M*. This should make *A* change its appearance in the business user's interface ⁴ to ease the inspection of motivation *M*. (ii) **Prefill** with content or **preselect answers** for the current question, so that the business user does not need to waste time in selecting/writing obvious answers. (iii) **Inspect all the answers** provided so far by the business user for the current questionnaire, to gather contextual information, and make technical decisions accordingly. (iv) **Alter the admissible answers** for the current question by adding or removing some of them from the current questionnaire. Novel answers may include *dynamically*-generated text, possibly specific to the current business user's situation, which could not be foreseen by the experts when designing the Q/A graph. (v) **Decorate the current question/answer** with additional details, such as charts, tables, textual explanations, etc, computed dynamically. (vi) **Read/write custom data** into a questionnaire-specific database, to store information to be retrieved later. By combining these operations, technical users can support business users in making informed decisions, in a timely fashion.

Discussion. FAIRBRIDGE leverages these mechanisms to enable (and, simultaneously, constrain) the fruitful collaboration between business users and technical users. The role of experts is crucial to designing the overall workflow the business and technical users should follow. Decisions are then taken by the business users, in collaboration with the technical users, while their interaction is mediated, tracked, and guided by FAIRBRIDGE.

Dually, the Q/A graph is a form of representation

⁴By being highlighted in green or red

of the FbD methodology. Updating the Q/A graph is equivalent to updating the methodology, and this in turn may step through adding more questions and answers to the graph, or changing their ordering. In any case, once a new version of the Q/A graph is ready, novel questionnaires being provided to business users will be based on the new version, hence putting a new version of the FbD methodology into practice.

4. Eliciting Non-technical Requirements

Social, ethical and legal experts are tasked with collecting the requirements from their respective fields by 1) analysing the domain on which the AI and identifying the issues that need to be taken into account and by 2) relying on the standard methods from their field. These actions (requirement identification and question design) are conducted *jointly* by the various experts. Experts should interpret the current state of the art in their field, and translate their *requirements* into a coherent set of questions and answers aimed at guiding business users' decision-making.

Legal requirements. The legal underpinnings of fair AI might come from different sources, such as the AI Act (Madiega, 2021) and the Charter of Fundamental Rights (Peers et al., 2021) Legislation is composed of many articles but tackling them simultaneously can be unwieldy. Legal experts should go through the legislation article by article and extract the key requisites from each article, translating them into questions and actionable items. We do not claim that the questions formulated in our meta-methodology guarantee overall legal compliance. However, we reckon that this is a reasonable trade-off, since considering the entirety of the legislation all at once can be an overwhelming task for even a large group of experts. We propose a leaner and more pragmatic approach that can be initiated by looking at individual articles and that can be iteratively refined by adding new requirements at later times.

Social Requirements. While legal requirements can be extracted by referring to the growing body of fair AI regulations, for sociologists, the process is less straightforward (Weinberg, 2022), and requires the involvement of the many stakeholders stacked along the AI lifecycle. Qualitative and participatory techniques are crucial to collect the very social needs underlying the current structure of social and institutional inequalities. These could then be translated into the requirements that computer scientists need to deploy into SW systems.

The most credited approaches for this involvement are: (i) participatory living experiments involving participants from different social and cultural backgrounds that are either under-represented or

Questions	Answers
Q1: In which situation are you among the following ones?	- Q1-A1: Data and algorithms available - Q1-A2: Only data available - Q1-A3: No data nor algorithms available
Q2: Provide the URL of the dataset on the Web or upload it directly	Free text answer
Q3: Indicate any sensitive attribute present in the data	List of answers computed by the technical user
Q4: What is your target fairness notion?	- Q4-A1: Calibration - Q4-A2: Independence - Q4-A3: Separation - Q4-A4 :Causality
Q5: What is your target fairness metric?	- Q5-A1: Statistical parity - Q5-A2: Equalised odds - Q5-A3: Equal opportunity - Q5-A4: Darlington criterion

Table 1. FairBridge in action. This is a snapshot of possible questions (Qx) and related answers (Qx-Ay).

more at risk of algorithmic discrimination; (ii) surveys to be submitted to the aforementioned affected social groups; (iii) focus groups to directly involve all the relevant stakeholders in the AI lifecycle. In essence, we open up our meta-methodology to limit and mitigate potential biases rising from the unequal bundle of social, economic, and institutional structures that characterizes the specific domain where the AI system will be applied. This is a long path to undertake, but it is a promising one to govern AI in the public interest.

Requirements Translation. Social and legal requirements must also be translated into quantifiable measures and methods. The requirements are expressed as a series of questions and allowed answers. Technical experts then examine them and propose technical measures to complement the questions, suggest modifications and add new ones. Questions and answers are represented in a shared, common format; and used to populate the FAIRBRIDGE’s Q/A graph.

5. Case Study

To exemplify the benefits of our approach, we present a use case where a branch of the Q/A graph is designed. The use case is deliberately concise, simplified, and incomplete, as it is meant to be illustrative rather than exhaustive. The questions and admissible answers are listed in Section 4, while Figure 3 and 4 display, respectively, the flow of the questionnaire and the sequence diagram.

Let us assume the root question in our Q/A graph is: **Q1**) with its admissible answers; **Q1** is single-choice. For the time being, we assume that experts want to focus on the case where answer **Q1-A2** is selected⁵. When business users have some data available but no trained algorithm yet, the purpose of the questionnaire is to guide them towards the detection and mitigation of any bias affecting the data, before training any AI algorithm. While the goal is straightforward, pursuing

⁵Other cases are left for future revisions of the Q/A graph

it implies stepping through a series of decisions, which in turn depends on the socio-legal context in which the AI system is going to be deployed, as well as on the technical details of the available data. Intuitively, detection should occur before mitigation and (i) it involves choosing one or more fairness metrics to be computed on the data. This, in turns, (ii) requires choosing a fairness *notion* to be targeted, – as each notion recommends specific metrics –, and (iii) it also requires identifying a set of *sensitive attributes* to be considered in the metrics computation. Finally, these attributes (iv) need to be selected from the available data. Accordingly, subsequent questions are designed to guide business users through these decisions.

Q2) admissible answer is either a free-text field or a file upload one. In this way, business users can provide the dataset they want to analyze, allowing for subsequent questions to be tailored to that data. **Q3)** has no admissible answers, as the experts cannot forecast which attributes will be present in the data of any particular business user. Instead, this questionnaire-specific answers for **Q3** are assumed to be computed by a technical user (see the * symbol in Fig.2), using the dataset provided by the business user in **Q2** as a source of information. In this way, when the business user will visualise **Q3**, they will find the names of the attributes of the dataset they provided among the admissible answers, with an indication of which are the sensitive ones, and why. To make the population of **Q3** possible, the technical user should write scripts reacting to both the ‘Answer given’ event of **Q2**, and the ‘Question asked’ event of **Q3**, and inspecting the dataset provided by the business user in response to **Q2**, to detect its attributes, and, among them, tagging the sensitive ones is recommended.

The next step (**Q4**) involves the selection of a fairness notion, currently allowing a selection among the most popular in the literature. The choice of the fairness *metrics* follows a similar pattern, as can be observed in **Q5**), again with admissible answers being the most

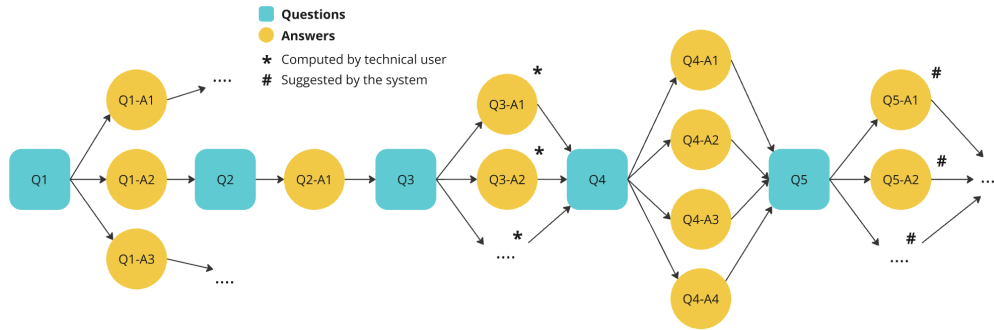


Figure 3. Questionnaire flow of illustrated use case. See Section 4 for the full questions and answers.

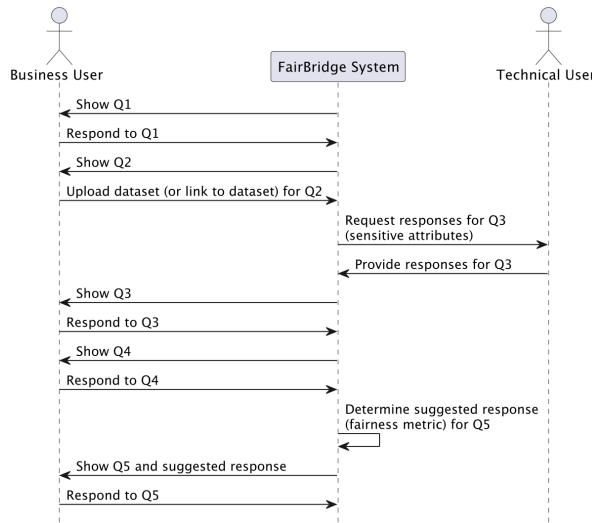


Figure 4. Sequence diagram of illustrated use case.

common fairness metrics from the literature —and any variation of them. To account for the fact that the choice of the fairness *metric* depends on the choice of the fairness *notion*, we assume the existence of one more script, reacting to the ‘Answer given’ event of **Q4** and suggesting the admissible answers for **Q5** accordingly.

At this point, FAIRBRIDGE has all the information it needs to compute whether, and w.r.t. which sensitive attributes, the dataset provided by the business user is biased, and contextually let the business user choose which cases require mitigation. **Q6** allows multiple-choice, with available answers including one entry for each sensitive attribute selected in **Q3**, and for each fairness metric selected in **Q5**. None of these entries could be foreseen by the experts when designing the Q/A graph, as they depend on the data provided by the business user. Thus, all admissible answers should be dynamically computed by an automated script.

Subsequent questions in the Q/A graph would involve the selection of a mitigation algorithm to be

applied to the data, and, later on, the training of a fair AI model on the data.

Discussion

The use case described so far is already enough to discuss the potential of our approach. Other than satisfying the principles discussed in Section 3, our approach comes with two major benefits that become evident in the long run: extensibility and automation.

Extending the Q/A Graph. Technically speaking, the Q/A is extended by adding new questions and answers to the graph, possibly modifying and re-disposing the existing ones. Most commonly, experts will add questions in groups, i.e., internally coherent branches of questions and answers to be attached somewhere in the graph. This allows experts to focus on *partial*, purpose-specific workflows, without worrying about the whole graph at once. For example, in the use case above, we only describe one possible branch of questions aimed at detecting biases in the data. Further branches could be added to the graph, e.g., to guide users towards the fair provisioning of data. This would be yet another partial workflow, branching from **Q1-A3** and eventually joining **Q2** once data has been collected.

About automation. While most activities provided by technical users are organization-specific, some of them are general enough to be automated at the system level. For instance, the computation of fairness metrics and the detection of biases in the data could be automated by FAIRBRIDGE directly, without the need for any particular technical user to reinvent the wheel.

We expect that this kind of automation will become more and more frequent as AI-fairness-related practices become more established. FAIRBRIDGE is designed to support this evolution, through its inherent event-driven architecture. Whenever a reaction to some ‘Question showed’ or ‘Answer given’ event is general enough, it can be implemented as a system-level script, and reused across different organizations. This could be for

instance the case, of the scripts described in the use case above, which are general enough to be included in FAIRBRIDGE since its design.

Generally speaking, automation scripts are the way by which technical aspects of the FbD methodology are injected into the design flow. These could be provided by FAIRBRIDGE itself, or by organizations' technical users, or even by third-party developers. This is another key benefit of our approach, as it allows for FAIRBRIDGE to evolve its degree of automation over time. While technical aspects may be initially delegated to technical users in the earliest stages of the methodology/system, they may be progressively automated as the methodology/system matures.

6. Conclusion

We address the problem of building FbD methodologies merging social, legal, ethical, and technical perspectives. This is a crucial step towards fair and unbiased AI. Our approach does not directly provide a methodology, as the varying social and legal factors require different strategies. Instead, we proposed a *meta-methodology* to build FbD methodologies. Our meta-methodology is called FAIRBRIDGE and is rooted in questions and answers, to be tackled by social, legal, and technical experts. We propose a structured flow of questions connected to admissible answers. We demonstrate its feasibility by illustrating a case study where questions and answers are organized in a graph structure (that allows for a straightforward algorithmic implementation). In future works, we plan to fully implement FAIRBRIDGE. Our approach will not substitute human decision-making but it will rather ease the task of the team of experts designing and developing fair AI applications.

Acknowledgements The work has been partially supported by the AEQUITAS project funded by the European Union's Horizon Europe Programme (Grant Agreement No. 101070363); and by the FAIR foundation, funded by the European Commission under the NextGenerationEU programme (PNRR, M4C2, Investimento 1.3, Partenariato Esteso PE00000013, Spoke 8 "Pervasive AI").

References

Calegari, R., Castañé, G. G., Milano, M., & O'Sullivan, B. (2023). Assessing and enforcing fairness in the ai lifecycle.

- Caton, S., & Haas, C. (2024). Fairness in machine learning: A survey. *ACM Computing Surveys*, 56(7), 1–38.
- de Almeida, P. G. R., dos Santos, C. D., & Farias, J. S. (2021). Artificial intelligence regulation: A framework for governance. *Ethics and Information Technology*, 23(3), 505–525.
- Franklin, J. S., Bhanot, K., Ghalwash, M., Bennett, K. P., McCusker, J., & McGuinness, D. L. (2022). An ontology for fairness metrics. *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 265–275.
- Madiega, T. (2021). Artificial intelligence act. *European Parliament: European Parliamentary Research Service*.
- Makhlouf, K., Zhioua, S., & Palamidessi, C. (2021). Machine learning fairness notions: Bridging the gap with real-world applications. *Information Processing & Management*, 58(5), 102642.
- Manjarrés, Á., Fernández-Aller, C., López-Sánchez, M., Rodríguez-Aguilar, J. A., & Castañer, M. S. (2021). Artificial intelligence for a fair, just, and equitable world. *IEEE Technology and Society Magazine*, 40(1), 19–24.
- Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6), 1–35.
- Mitchell, S., Potash, E., Barocas, S., D'Amour, A., & Lum, K. (2021). Algorithmic fairness: Choices, assumptions, and definitions. *Annual review of statistics and its application*, 8, 141–163.
- Peers, S., Hervey, T., Kenner, J., & Ward, A. (2021). *The eu charter of fundamental rights: A commentary*. Bloomsbury Publishing.
- Shneiderman, B. (2020). Bridging the gap between ethics and practice: Guidelines for reliable, safe, and trustworthy human-centered ai systems. *ACM Transactions on Interactive Intelligent Systems (TiIS)*, 10(4), 1–31.
- Wachter, S., Mittelstadt, B., & Russell, C. (2021). Why fairness cannot be automated: Bridging the gap between eu non-discrimination law and ai. *Computer Law & Security Review*, 41, 105567.
- Weinberg, L. (2022). Rethinking fairness: An interdisciplinary survey of critiques of hegemonic ml fairness approaches. *Journal of Artificial Intelligence Research*, 74, 75–109.