

DELFT UNIVERSITY OF TECHNOLOGY

MASTERS THESIS

---

**Multi-Level Fairness Framework : A  
Socio-Technical framework for Fairness  
Requirements Engineering in Machine  
Learning**

---

*Author:*

Manisha Sethia

*Thesis Committee:*

*Chair:*

Prof. Dr. A. van Deursen, Faculty EEMCS, TU Delft

*University Supervisor:*

Dr. C. Lofi (Web Information Systems, EEMCS, TU Delft), A. Balayn ((Web Information Systems, EEMCS, TU Delft)

*Company Supervisor:*

Dr. Rüya Gökhan Koçer, ING Bank Supervisor, Dr. Flavia Barsotti, ING Bank Supervisor

*A thesis submitted in fulfillment of the requirements  
for the degree of Master of Science*

*in the*


**Web Information Systems Group  
Software Technology**

September 22, 2021

## Declaration of Authorship

I, Manisha Sethia, declare that this thesis titled,

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: 

---

Date: 22/09/2021

---



*“I know the world isn’t fair, but why isn’t it ever unfair in my favor?” –Bill Watterson ”*



DELFT UNIVERSITY OF TECHNOLOGY

# *Abstract*

Electrical Engineering, Mathematics and Computer Science  
Software Technology

Master of Science

## **Multi-Level Fairness Framework : A Socio-Technical framework for Fairness Requirements Engineering in Machine Learning**

by Manisha Sethia

Machine Learning models are begin increasingly used within the industry such as by financial institutions, governments and commercial companies. In the past few years, there have been several incidents where these ML models show discriminatory behavior towards particular groups of people, leading to unfair decisions that can have negative impacts on the lives of these people. Therefore, eliminating bias and ensuring fairness within these models is crucial to the societal expectations these institutions would like to meet.

While there are tools and research towards technical mitigation methods for bias and unfairness. There is a lack of focus on the process of implementing these tools and methods within the industry, to develop ML models with the consideration of fairness at an early stage. In particular, there is a lack of specification on what fairness goals and objectives the ML model should accomplish.

Without having this clarity, industry stakeholders can apply the tools and algorithmic unfairness mitigation methods but if the fairness requirements are not defined, then a) one is still not sure whether the ML model is solving the correct fairness goals and b) one is still not sure what the trade-offs and feasibility within different fairness goals, and available resources may look like. To address this, we design a Multi-Level Fairness Framework (digital workflow) that aims towards supporting stakeholders within the industry to perform Requirements Engineering, specifically elicitation and modeling, for fairness in a Machine Learning Model. Firstly, we gather practices within research that can aid in performing Fairness Requirements Elicitation and Modeling (F.R.E.M). Secondly, we investigate the industry challenges for conducting F.R.E.M. via conducting a qualitative study with nine interviews within ING Bank N.V. (a participatory financial institution) across three ML models. We discover that a possible solution should target towards three challenges, namely raising consideration on aspects of fairness, facilitating specification for defining fairness and aiding communication by displaying information and terminology understandable to the background of the various stakeholders involved.

We, then reflect on our findings to specify requirements for the framework and the design for the Multi-Level Fairness Framework. We evaluate the framework, by building a digital prototype of the framework, and conducting a qualitative study consisting of four interviews within ING and investigate the effects of the framework in targeting the challenges of consideration, specification and communication to support stakeholders in F.R.E.M.

Fairness, Machine Learning, Requirements Engineering, Framework, Industry Practices, Binary Classification





## *Acknowledgements*

This thesis was a learning experience that challenged me in various ways, both intellectually and personally. I am grateful to be part of, and make my contributions to the research field that addresses fairness within my discipline, and strives towards achieving this in practice.

From the very start of this project, there have been notable people that have guided me, given me their time and efforts on conducting this project. I would like to thank my university supervisors, Christoph Lofi and Agathe Balayn for their critical point of view, support, time and guidance throughout the project. I would like to express my gratitude towards the AIForFintech Lab, the people that make-up that lab, especially Arie van Deursen, Elvan Kula and Luis Cruz, and inspire research to be conducted within industry. I would like to thank my company supervisors from ING Bank, Rüya Gökhan Koçer, and Flavia Barsotti for their guidance, viewpoints, time and organizing with me to conduct my research experiments. To all the participants that were involved in this project, I would thank you for your interest and enthusiasm. All these people have played a crucial role in this thesis, and I can only be grateful for such support.

Last but not least, I would like to thank my family, especially my parents for being there every step of the way to support me, and encourage me. And my friends, for supporting me, and being there for me.



# Contents

<b>Declaration of Authorship</b>	<b>iii</b>
<b>Abstract</b>	<b>v</b>
<b>Acknowledgements</b>	<b>ix</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Problem Statement . . . . .	2
1.2 Scope of the project . . . . .	4
1.3 Research Questions, Contributions and Challenges . . . . .	5
1.4 Organization . . . . .	7
1.4.1 ING Bank N.V. . . . .	8
<b>2 Background and Related Work</b>	<b>9</b>
2.1 Related Work . . . . .	9
2.2 State-of-the-art in eliciting fairness requirements . . . . .	10
Takeaways . . . . .	13
2.3 State-of-the-art in specifying fairness requirements . . . . .	13
Takeaways . . . . .	17
2.4 State-of-the-art in modeling fairness requirements . . . . .	17
Take-aways . . . . .	19
Conclusion . . . . .	20
<b>3 Framework Requirements Elicitation and Specification</b>	<b>23</b>
3.1 Institutional Challenges . . . . .	24
3.1.1 Stakeholder Study . . . . .	26
3.2 Requirements for the framework . . . . .	34
3.3 Insights for the framework . . . . .	39
<b>4 Multi-Level Fairness Framework</b>	<b>43</b>
4.1 Objectives . . . . .	43
4.2 Components of the framework . . . . .	44
4.3 Actions for the framework . . . . .	48
4.4 Mechanisms for the M.L.F.F. . . . .	48
4.5 Characteristics of the framework . . . . .	50
4.5.1 Multi-Level Nature . . . . .	51
4.5.2 Information Nature . . . . .	51
4.5.3 Role-Based Nature . . . . .	52
4.5.4 Challenge-Nature . . . . .	56
<b>5 Evaluation</b>	<b>59</b>
5.1 Objectives . . . . .	59
5.2 Methodology . . . . .	60
5.3 Results . . . . .	63

5.3.1	Analysis	63
5.3.2	Insights	65
<b>6</b>	<b>Conclusion</b>	<b>69</b>
6.1	Reflection	69
6.2	Conclusion	73
<b>A</b>	<b>Appendix A</b>	<b>75</b>
<b>B</b>	<b>Appendix B</b>	<b>81</b>
B.1	Stakeholder Study A	81
B.1.1	Model Owner	81
B.1.2	Model Developer	82
B.1.3	Domain Expert	83
B.2	Stakeholder Study B	85
B.2.1	Model Owner	85
B.2.2	Model Developer	86
B.2.3	Domain Expert	88
<b>C</b>	<b>Appendix C</b>	<b>91</b>
C.1	Evaluation Stakeholder Study in Digital Prototype	91
C.1.1	Introduction and Common Components	91
C.1.2	Model Developer	91
C.1.3	Model Owner	91
	<b>Bibliography</b>	<b>99</b>

# List of Figures

2.1	Confusion matrix of statistical notions [77]	15
2.2	High Level Overview of Findings from Background Study	21
3.1	Roles within Stakeholder Study for ML Model Development	26
3.2	Overview of different levels of knowledge, accessed by different methods [66]	27
3.3	Status Quo Procedure for Communication between stakeholders	30
3.4	Status Quo Procedure for Fairness Responsibilities between stakeholders	31
4.1	Framework with Sub-Mechanisms	50
4.2	Multi-Level Fairness Framework	51
4.3	Model Owner Engagement with the Framework	54
4.4	Model Owner Engagement with the Framework	55
4.5	Model Owner Engagement with the Framework	57
C.1	Introduction Text	92
C.2	Component : Model	92
C.3	Selecting Role	92
C.4	Component : Case Studies	93
C.5	Component : Dilemmas	93
C.6	Component : Social Fairness Notions - Reviewing inequalities	94
C.7	Component : Fairness Definitions - Reviewing fairness definitions	94
C.8	Component : Fairness Metrics	94
C.9	Component : Technical Mitigation Methods	95
C.10	Component : ML Pipeline Stages	95
C.11	Component : ML Pipeline Activities - Adding an activity to a stage	95
C.12	Component : ML Pipeline Stages and Activities - Reviewing the ML Pipeline	96
C.13	Component : ML Pipeline Stages and Activities - Adding a fairness step to the ML Pipeline	96
C.14	Component : Trade-offs	97
C.15	Component : Social Fairness Notions - Specifying inequalities	97
C.16	Component : Social Fairness Notions - Reviewing fairness philosophies (generated from mappings)	98
C.17	Component : Fairness Definitions - Specifying fairness definitions	98



# List of Tables

1.1	Organization of this project	8
2.1	Machine Learning Model - Stages and Example Activities	20
3.1	Stakeholder Study Setup	29
3.2	Summary of Stakeholder Study Mapping to Objectives	29
3.3	F.R.E.M Elicitation, and Modeling Notations noted from [17]	35
3.4	Requirements for designing the framework	39
3.5	F.R.E.M Elicitation, and Modeling Methodologies, Strategy and Advice noted from [17]	40
3.6	Insights for designing the framework	42
4.1	Targeted Challenges of the components	58
5.1	Stakeholder Study Setup	60
5.2	Primary, Reference and Repetitive Tasks	64
A.1	Levels of inequalities [46]	75
A.2	Fairness Philosophies and corresponding inequalities mentioned in [46]	76
A.3	Fairness Dimensions (Procedural and Distributive) specified by [45]	76
A.4	Fairness Techniques (for Binary Classification) mentioned in [15]	77
B.1	Stakeholder Study Mapping to Objectives	90





# List of Abbreviations

- F.R.E.M.** Fairness Requirements Elicitation and Modeling (for this project)  
**M.L.F.F.** Multi-Level Fairness Framework



*For/Dedicated to/To my parents, Jai Sethia and Pooja Sethia...*



## Chapter 1

# Introduction

As we move towards a world where industries focus on data-driven solutions and continue to embrace the power of automation from machine learning, we also move towards a world which brings many new, unfamiliar challenges that can have a significant effect on the people and world around us. For example, in banks, whether or not the end-user is given a loan, can be dependent on a decision-making ML algorithm. In recruitment systems, potential employees are screened and accepted and rejected by a ML model, before even having the opportunity to interview. As such, a lot of these ML-driven solutions can have severe impacts on the end-user's life.

Machine learning model outcomes can be misaligned to the intention of its creators and societal expectation, such as discriminating based on personal demographic attributes, e.g. gender and race [46]. As such, machine learning (ML) systems can exhibit or even amplify social inequities and unfairness [38].

What do we mean by social inequities and unfairness? That in itself is a debated topic within ethics, sociology and philosophy. In practice, the institution building and deploying the ML model, has its set of societal expectations to meet, in terms of ensuring fairness. These expectations can be legal, policies or ideologies that the institution wants to abide by.

To demonstrate mismatches and dangers of deploying ML models within industry, we highlight past incidents :

### 1. Classification Model within the criminal justice system

Correctional Offender Management Profiling for Alternative Sanction (COMPAS) is a risk recidivism (tendency of a criminal to re-offend) assessment system that was in use within US Courts for pre-trial detention. ProPublica reporters discovered that the model was discriminating against black people, as amongst the offenders that did not re-offend in two years, the COMPAS model had consistently assigned black people worst scores that white people [41].

### 2. Image Recognition System within search engines

In 2015, there was an incident of Google Photos labeling black people as *gorillas*, which cause a media public outrage. In 2018, [14] found that the rate for misclassification of black women within a gender classification software was higher than other groups of varying gender and race.

### 3. Recommendation Systems within search engines

[43] found a systemic un-representation of women of various occupations, in image search engines. In 2013, Google was found to deliver ads suggestive of arrests more to black people than to white people in it's ad delivery system, AdSense [].

To tackle discrimination and unfairness, research towards fairness-aware machine learning systems has been conducted. Over the last few years, researchers have developed fairness definitions, (un)fairness mitigation methods and tool-kits for machine learning algorithms to investigate unfairness (e.g. AI Fairness 360 by [3], What-If by [80], Aequitos [1]) towards fair machine learning models.

When considering these machine learning models being built within the industry, fairness is then no longer limited to a simple application of an algorithmic method, or exploration via one of these tools but rather a process that can address meeting the societal expectation that the industry wants to achieve or maintain. For example, [39] investigates the challenges from the Machine Learning practitioners' within industries' perspectives. One challenge listed is that there is "lack of tools and process to support practitioners in identifying the components in ML where the bias issues occur. For example, should efforts be focused on training data or on the model itself". Another challenge mentioned is that the humans developing the models may introduce bias into the models, i.e. "bias in the human loops". [34] also identify pragmatic challenges of addressing fairness within the industry, such as prioritization of correcting bias, proposing minimum viable products and addressing technical debts in cultural change. Prioritization of correcting bias suffers due to engineering teams being focused on a carefully planned road map for product delivery, maintenance and improvement. This can surface pressing priorities that compete with priorities of addressing fairness within these products. Proposing minimum viable products refers to the adoption of agile processes and iterative nature of building a product, and inability to address fairness at once within the product. Addressing technical debts in cultural change indicates the need for longer term cultural change and education toward bias-awareness.

These works indicate that industry settings bring on a new set of challenges on top of existing fairness in ML issues, for example addressing entirety of the ML model development pipeline or dealing with industry practices in ML model development (i.e. agile processes).

So, despite technical mitigation methods and tools being available publicly, questions on fairness within an ML model for institutions within the industry still linger. For example, what is considered to be fair/unfair within a ML model, to what extent do they ensure it, what steps need to be taken within the Machine Learning pipeline to ensure these fairness goals, to what extent are the fairness goals realizable in terms of resources?

[38], [28], and [34] highlight that there is an urgent need for internal processes and tools to support companies in developing fairer systems in the first place.

[34] concludes that combining the various fields of research within fairness and machine learning to create an internal understandable framework for industry teams is most fruitful in ensuring that ML models can be developed within fairness in mind, at an early stage.

## 1.1 Problem Statement

The question that follows is, what exactly should this internal understandable framework, or internal processes address? In software engineering within industry, requirements engineering is an important phase used to translate the imprecise, incomplete needs and wishes of the potential users of software into complete, precise and formal specifications. Requirements engineering, not only helps in building the correct software, but also allows for management of cost, resources and time, at an

early stage of the model development process. Similarly, we envision that requirements engineering for fairness can benefit industries to understand the fairness requirements to be addressed in Machine Learning models, and manage the trade-offs that occur with these requirements. It can also provide the model development team on guidance on what fairness means for the model, and support in understanding what the tools and mitigation methods should be accomplishing. [30] mentions that most ML models' unfairness stems from a unclear specification of fairness requirements to accomplish, test and validate.

We envision that creating a framework for supporting fairness requirements engineering within an institutional setting as a step towards developing fairer ML models. Here, we note that simply applying requirements engineering processes present within the industry setting to generate fairness requirements does not necessarily address the nature of fairness requirements. Fairness requirements are **subjective, uncertain** and **variable**. **Subjective** can make fairness requirements be contradictory, multi-objective and sometimes unattainable and so elicitation and modeling of these requirements may involve trade-offs to be made, realizability to be recognized and prioritization to be placed [9], [79], [42]. **Uncertainty** brings an *iterative* nature to the specification of these fairness requirements wherein requirements may be recognized only after the model is deployed, or become more concise over time [63], [53], [26]. **Variable**, in the sense, that fairness requirements may differ from situation to situation, thus require contextual considerations and changes over time to be accommodated [13].

In this thesis, we aim to design and evaluate a (digital) framework which supports the industry stakeholders in performing fairness requirements engineering for ML models. We note that :

1. Fairness requirements engineering is scoped to include fairness requirements elicitation and fairness requirements modeling, which we will refer to as F.R.E.M. Fairness Requirements elicitation refers to exploring and defining fairness goals, objectives and motives for a particular ML model. Fairness Requirements modeling refers to understanding the entities, behaviors and constraints of the fairness goals, objectives and motives in relation to the ML model (for this thesis, the framework should support modeling the feasibility and trade-offs between fairness goals, objectives and the ML model)
2. By framework, we mean a (interactive, digital) workflow that guides the stakeholders to perform F.R.E.M. For instance, for fairness requirements modeling, the framework is not aimed towards determining feasibility, but rather providing prompts and guidance to the stakeholders such that they can determine the feasibility. Similarly, for fairness requirements elicitation, the stakeholders cannot determine what fairness means directly from the framework, but rather use the framework to think around what fairness goals, objectives and motives they would like to determine. **In other words, the framework is a workflow/guide for performing F.R.E.M, rather than a tool to perform F.R.E.M.**
3. By industry stakeholders, we mean that the framework is designed for the stakeholders building ML models within a industry setting, addressing numerous challenges associated with fairness requirements engineering within the industry and addressing a typical ML model development set up within the industry.

## 1.2 Scope of the project

We first introduce the reader to the definition of requirements engineering itself. Next, we want to be aware of what we can expect from a requirements engineering solution. We want to investigate what the nature of this can look like, so we can be considerate of this while designing our solution. This can also help us eliminate any misconceptions on the nature of requirements engineering, that one may have.

Then, we scope the stages in requirements engineering to our project. In other words, we outline which primary stages we for-see our framework delivering.

### What is Requirements Engineering?

[52] describes Requirements Engineering as follows : "Broadly speaking, software systems requirements engineering (RE) is the process of discovering that purpose, by identifying stakeholders and their needs, and documenting these in a form that is amenable to analysis, communication, and subsequent implementation.

Requirements engineering is the branch of software engineering concerned with the real-world goals for, functions of, and constraints on software systems. It is also concerned with the relationship of these factors to precise specifications of software behavior, and to their evolution over time and across software families."

### Stages of Requirements Engineering

[17] provide a recent and comprehensive outline of the stages involved in Requirements Engineering through their survey analysis. We present the requirements engineering stages :

1. **Requirements Elicitation** : Requirements elicitation comprises activities that enable the understanding of the goals, objectives, and motives for building a proposed software system. Elicitation also involves identifying the requirements that the resulting system must satisfy in order to achieve these goals.
2. **Requirements Modeling** : In requirements modeling, a project's requirements or specification is expressed in terms of one or more modeling notations. Modeling notations help to raise the level of abstraction in requirements descriptions by providing a vocabulary and structural rules that more closely match – better than natural language does – the entities, relationships, behavior, and constraints of the problem being modeled.
3. **Requirements Analysis**: Requirements analysis assesses the quality of requirements models and documentation.
4. **Validation and Verification**: Requirements validation ensures that models and documentation accurately express the stakeholders' needs. In cases where a formal description of the stakeholders' requirements exists, obtained perhaps by validation, verification techniques can be used to prove that the software specification meets these requirements.

From the stages of requirements engineering, we see that **Requirements Elicitation** and **Requirements Modeling** are targeted towards the formulation of requirements. **Requirements Analysis** and **Requirements Validation and Verification** are targeted towards the quality and testing of the formulated requirements.



As we wish to provide a framework pertaining to the formulation of fairness requirements, the relevant stages for our framework will be **Requirements Elicitation** and **Requirements Modeling**.

*Remark.* We realize that **Requirements Analysis** and **Requirements Validation and Verification** play a critical role in facilitating re-formulation. This can be a possibility for future work to look into.

As limited work has been done within fairness requirements engineering, there is no definition coined for requirements elicitation and modeling for fairness. One should note that we understand that requirements specification to be incorporated within elicitation and modeling. For example, specifications are defined after the elicitation phase, but re-specification may be needed after the modeling phase.

### 1.3 Research Questions, Contributions and Challenges

In this section, we outline how we tackle these research questions by providing the corresponding contribution for each research question. Furthermore, we introduce the reader to challenges we for-see, in this project.

#### Research Questions and Contributions

Let us consider the problem statement :

"Our aim is to design and evaluate a framework that supports fairness requirements elicitation and modeling, in an institutional setting."

We can break this down into four research questions we want to investigate :

1. **Research Question 1** What are the state-of-the-art practices to support fairness requirements elicitation and modeling?

In this question, our main hurdle is the limited work present in addressing fairness requirements engineering. This means, we need to perform several literature studies and extrapolate which methods/practices may actually aid in fairness elicitation, and modeling.

Here, the contribution is an overview of various methods/practices that can be connected to perform fairness elicitation and modeling.

(a) **State-of-the-art in eliciting fairness requirements**

We want to discover the state-of-the-art fairness concepts that aid in deriving the suitable fairness definitions for Machine Learning model. The takeaway is an overview of methods that aid in investigating fairness goals, objectives and motives.

(b) **State-of-the-art in specifying fairness requirements**

We want to investigate the state-of-the-art fairness definitions that can be implemented empirically in a Machine Learning model. The takeaway is an overview of methods that can be included within the framework to allow for specifying fairness goals, objectives and motives, in a formal and empirical manner.

(c) **State-of-the-art in modeling fairness requirements**

We want to investigate the state-of-the-art methods that allow for the trade-offs and feasibility to be captured for fairness requirements. The takeaway is an overview of methods that can be included within the

framework to prompt investigating trade-offs and feasibility of specified fairness requirements.

(d) **State-of-the-art practices in requirements elicitation and modeling**

We want to investigate the state-of-the-art tasks required in requirements engineering and modeling. The takeaway is an overview of practices for requirements elicitation and modeling that we can later adapt to address the fairness institutional challenges and include within the framework.

**Contribution :** A overview of methods, practices and tasks that can be relevant to consider in eliciting, modeling (including specifying) fairness requirements.

(a) **Research Question 2** What are the institutional challenges when conducting fairness requirements elicitation and modeling?

(b) **Institutional Challenges**

Using the case of a the participatory institution (i.e. ING Bank N.V.), what are challenges to address in order to support fairness requirements elicitation and modeling within an institutional setting?

**Contribution :** A study of institutional challenges to address when engineering fairness requirements, specifically eliciting, modeling (including specifying) fairness requirements.

By reflecting on Research Question 1 and Research Question 2, we can understand the steps and techniques required to support fairness requirements elicitation and modeling within the institutional setting. Based on this, we can formulate requirements for the framework and design the framework itself, and evaluate the framework. This leads to the following Research Questions :

2. **Research Question 3** What are the requirements and design of the framework?

(a) **Requirements of the framework**

We reflect on the overview of tasks in requirements elicitation and modeling found in Research Question 1, to adapt to specifying tasks that address the institutional challenges of supporting fairness requirements engineering. We will refer to these tasks as the requirements of the framework.

(b) **Design of the framework**

We reflect on the requirements specified for the framework and see which methods in state-of-the-art fairness requirements elicitation, specification and modeling found in Research Question 1, can be included within the framework. This provides us with the base components, mechanisms and actions that make-up the framework.

**Contribution :** A framework aimed at supporting stakeholders involved in Machine Learning model development to elicit, model (including specify) fairness requirements for a ML model whilst addressing institutional challenges discovered in Research Question 2.

(a) **Research Question 4** To what extent does the framework support fairness requirements elicitation and modeling for the institutional challenges identified?

**Contribution :** A qualitative study on the effects of the framework in addressing the institutional challenges in fairness requirements elicitation, and modeling. This includes presenting any additional findings or intriguing observations found in engagement with the framework, and performing requirements elicitation and modeling.

## Challenges

The challenges we need to address in achieving the problem statement manifests in many directions. Firstly, fairness within Machine Learning models needs to take into consideration social fairness, empirical fairness, machine learning technicalities and stakeholders (and institutional procedures) of the technology to provide a solution for fairness requirements engineering. This means, we need to constantly address the solutions from multiple perspectives. This is crucial to address the subjectivity and *bureaucracy* and technical complexity that comes with fairness.

This also leads to the challenge of having to understand social aspects, technical aspects and institutional aspects wherein different people involved (stakeholders), with backgrounds need to be able to work with the same solution. A sub-challenge within this, is already identified the problems within accomplishing fairness within a machine learning model being used within the industry. We find limited work addressing the problems and status of current knowledge and procedures of fairness for machine learning within the industry. This means, not only do we have to design a solution, but first we need to investigate the gaps to solve with a closer lens.

Another challenge is addressing the emerging interest in fairness for Machine Learning research. There is new research being developed in all these different fields of fairness and Machine Learning. For example, combining social fairness for machine learning is being studied, new (un)fairness mitigation methods are being developed and new tangents of empirical fairness definitions or fairness metrics are being coined. We need to design a solution that can absorb new research in an efficient manner. This means, we need to put pressure on the adaptability, and expand ability of our solution so it stays relevant and accommodating.

We also aim to combine the field of software engineering practices to accommodate fairness for machine learning. The work within this field remains limited. This cross-combination of fields requires us to perform deep inferences from literature, to design a solution we can deem is *backed* by literature.

Lastly, streamlining the evaluation of systems that deal with social, technical and overall subjective manners can be tricky. We need to account for the qualitative nature of this, and design evaluations that are open to new insights being inferred. This can allow us to not only evaluate our solution, but present interesting insights that were gained from testing our solution. We deem this to be of importance within the fairness and ML field, as insights from different perspectives can be correlated and benefit the field all together.

## 1.4 Organization

In Table 1.1, we present the objectives, research questions and contributions with their corresponding Chapter and Section. This will provide the reader of a clear overview on the organization of this project report. We structure the design and evaluation of the framework, similar to an agile software development project, wherein we elicit and specify requirements for the framework. We build a prototype of the

Research Question	Chapter
Research Question 1	Background and Related Work
Research Question 2	Framework Requirements Elicitation and Specification
Research Question 3	Designing the Multi-Level Fairness Framework for F.R.E.M.
Research Question 4	Evaluation of the Multi-Level Fairness Framework for F.R.E.M.

TABLE 1.1: Organization of this project

framework and evaluate it, with the intention that the findings can facilitate future work on developing another iteration of the prototype, or bring to light new requirements that need to be specified.

Research Question 1 and Research Question 2 focus on eliciting requirements for the framework, so gaining understanding on F.R.E.M. within an institutional setting through literature and interviews. Research Question 3a then aims to reflect on the elicitation, and specify requirements and insights for the framework. Research Question 3b then reflects on the requirements and insights, to then design the framework. Research Question 4 focuses on the evaluation of the framework via a digital prototype and a critical discussion on the quality and content of the findings, and opportunities for future work.

#### 1.4.1 ING Bank N.V.

We utilize the context of a financial institution, namely, ING Bank N.V., to construct the framework. This financial institution consists of ML models being built for Fin-Tech purposes, that can range from credit -risk modeling, to in-house operations. Each of these models provide a different fairness considerations to take into account.

## Chapter 2

# Background and Related Work

Previously, we outlined the problem statement for the project, namely, providing a solution for Requirements Engineering for Fairness in Machine Learning. In this Chapter, we will investigate Research Question 1. This exploration extends beyond gathering related work as there is limited work present in fairness requirements engineering.

We take the following approach to address this, namely:

### 1. Objectives

#### (a) State-of-the-art in eliciting fairness requirements

In this chapter, we want to discover the state-of-the-art fairness concepts that aid in deriving the suitable fairness definitions for Machine Learning model.

#### (b) State-of-the-art in specifying fairness requirements

In this chapter, we want to investigate the state-of-the-art fairness definitions that can be implemented empirically in a Machine Learning model.

#### (c) State-of-the-art in modeling fairness requirements

In this chapter, we want to investigate the state-of-the-art methods that allow for the trade-offs and feasibility to be captured for fairness requirements.

2. **Methodology** : We study literature to explore each research question. We streamline our exploration by aiming towards finding methods that connect elicitation, specification and modeling stages in some manner. This adds utility to the methods explored as each method can then result in an outcome which can be Incorporated through the stages.

3. **Result** : A high-level overview of the methods explored and the connections between these methods, in facilitating fairness requirements elicitation, modeling and specification.

## 2.1 Related Work

Within the related work, we look for literature addressing fairness requirements engineering. While there is quite some work within assuring, or defining fairness, these works do not directly address fairness requirements engineering. We will explore these works further ahead in our background study. In this section, we specifically look for works that address fairness requirements engineering. We extend the exploration to involve fairness requirements engineering for software systems as well, due to the limited nature of fairness requirements engineering research.

We find that addressing fairness within machine learning systems, or even software systems occur by explicitly examining discrimination or emerging fairness requirements after implementation [8], [33], [53], [23]. [45] proposes a framework towards fairness requirements engineering for algorithmic decision making. The framework focuses on formalizing fairness goals, and objectives, but does not address the dynamic or institutional environment. One framework proposed to address defining fairness in software systems regardless of the particular kind of discrimination, and in a dynamic environment is [30]. The study proposes a adaptive fairness model, which is fairness requirements-driven, and resource-driven. The fairness requirements-driven, comes in the form of including fairness requirements model which encapsulates high level fairness requirements given by the stakeholder. The resource-driven comes from trade-offs relating to the operations of the software system. [30] introduces the concept of adaptive fairness, and that fairness requirements are changing and must to be adapted towards. We still find a lack of research on fairness requirements elicitation and modeling, in the context of technical trade-offs and feasibility, and within an industry setting. This prompts us to look further into literature, to see which practices, methods and ideas we can gather that can facilitate fairness requirements elicitation and modeling.

## 2.2 State-of-the-art in eliciting fairness requirements

In this section, we focus on fairness within sociology, philosophy and Machine Learning to understand how one can identify their fairness goals and objectives, and the fairness entities involved. We start by understanding what fairness definitions exist beyond the Machine Learning community.

Fairness has been studied by sociologists ([40]), philosophers, economics and Machine Learning. Fairness reflects objective features of how people share resources. However, fairness attribution varies between contexts and individuals: laborers' wages might not seem unfair considered alongside colleagues yet extremely unfair alongside executives' salaries. ([55]). We select sociology and philosophy, as these fields explore the meaning of fairness pertaining to societal expectations. Looking at these fields will ensure that we cover different definitions of fairness, even the ones that may not have been popularized in the field of Machine Learning until now.

In particular, for fairness within sociology and philosophy, we focus towards concepts that have been shown to affect algorithmic decisions in some way.

We start from abstractly exploring fairness and move towards the concrete definitions and empirical metrics supporting fairness within Machine Learning. Our aim for doing so, is two fold, namely : How to communicate goals and observations and How can we extract entities for fairness.

To understand the communication of goals and observations and extraction of fairness entities, one needs to understand the nuances between different fairness notions used within research. For example, suppose if we have two fairness definitions that are similar but differ on one criteria. To allow for communication of goals and objectives effectively, identifying this particular criteria becomes crucial. Similarly, to facilitate fairness entities interactions, we need to be able to extract exactly what is considered an entity, what types of entities can be there and how these types of entities can interact.

To that end, we start by build our understanding on Fairness within sociology, and philosophy (which we will refer to as Social Fairness) and then move towards

re-connoitring the manifestation of these philosophies within current research field of Fairness within Machine Learning.

This will allow us to construct an overview, between fairness philosophies and definitions and metrics used within the Fairness in ML field, to get an idea of the gaps and connections that exist when moving from abstract fairness notions, to the more concrete fairness definitions and metrics. We can use this overview to then outline how we for-see the the communication of goals and objectives and entity interactions in fairness.

### Fairness in Sociology

We look towards survey literature to understand the field of fairness within sociology to get as complete view as possible. [35] surveys the theories within organizational justice, a field belonging to sociology, to **taxonomize** theories aimed towards creating fairness. [35] groups theories into proactive content and processes.

Proactive content touches upon how to create fair allocation (for example, fair payment of laborers), whilst proactive processes are theories aiming towards creating fair procedures (for example, policies and procedures towards laborers are to be fair).

For example, one theory within this proactive content is studied and proposed by [35], [48], which propose that allocation of resources can be done in a equitable fashion, where they experiment to show that resource allocation can be done equally or in accordance to participant needs.

Another theory within this category is **Justice Motive Theory** which is based on four principles : (a) competition-allocations based on the outcome of performance, (b) parity-equal allocations, (c) equity- allocations based on relative contributions, and (d) Marxian justice-allocations based on needs.

Similarly, [35] taxonomizes proactive processes, which consist of theories that tackle fair processes. One theory considered a proactive process is **Allocation Preference Theory**, which states that "certain procedures will be differentially instrumental in meeting their goals, and that the procedure believed to be most likely to help attain one's goal will be the most preferred one."

This taxonomy of proactive content and process is reflected in [22] which mentions four dimensions in sociology, namely **Procedural fairness**, **Distributive fairness**, **Informational fairness** and **Interpersonal fairness**. Herein, **Distributive fairness** and **Procedural fairness** dimensions target whether the allocation is based on what a subject deserves (or needs) and whether decision-making processes are fair, respectively.

Now that we understand the segregation of the theories into the two taxonomies, we still need to identify how the multiplicity of proactive content (and process) (or procedural and distributive fairness) allocation strategies can look like.

For procedural fairness, **Allocation Preference Theory** provides eight principles that can guide this decision. These principles include that: (a) allow opportunities to select the decision-making agent, (b) follow consistent rules, (c) are based on accurate information, (d) identify the structure of decision-making power, (e) employ safeguards against bias, (f) allow for appeals to be heard, (g) provide opportunities for changes to be made in procedures, and (h) are based on prevailing moral and ethical standards. [35] shows the general support of studies that back these principles.

For distributive fairness, [22] means three principles namely outcomes are based on **equity**, **equality** and **need**.

### Fairness in Ethical and Political Philosophy

Looking into philosophy, ethical philosophy deals with the study of how (and if) equality should be pursued in society. Yet again, we flock towards survey literature to understand the field of fairness within ethical and political philosophy to strive towards exhaustiveness.

For example, egalitarianism refers to equality within society.

"Egalitarians have articulated various competing views, including welfare, understood in terms of pleasure or preference-satisfaction [21]; resources such as income and assets [57]. Others propose that inequalities in welfare, resources, or capabilities may be acceptable, so long as citizens have equal political and democratic status [59]."

[10] highlights the importance of egalitarianism and justice in the context of algorithmic fairness.

For example, let us take a loan approval process (which can be seen as a resource allocation process), the system is categorizing people into outcome classes, which then result in some form of a positive or negative effect. A negative effect, here, can be seen as being denied a loan. The differing perspective of egalitarianism come into play here, as for example, **luck egalitarian** aim of pursuing redistribution only where inequalities are due to pure luck. This means, that inequalities that are a result of luck, and not an informed choice or decision can be indeed considered an inequality. For example, **race** can be considered as luck, and any inequalities on loan allocation made on the basis of race should be redistributed. Furthermore, **luck egalitarianism** tolerates any conditions, for example, **neighborhood**, that may be arising from **race** to be tolerated as **luck**, and inequalities be addressed. In critical discussion of luck egalitarianism, [70] states that "luck egalitarianism should only be sensitive to responsibility for creating advantages and disadvantages – not to responsibility for distributing them". In other words, completely excluding informed choices as a reason for inequality may not be justified, if these informed choices are leading towards societal prosperity. For example, if a worker at an Non-Governmental Organization works towards a social cause, at a voluntary lower income, then the denial of a loan can be treated as an inequality and reconsidered.

[46] finds a similar identification of ethical philosophies defining **what is equality**, and provides layers of inequality to be considered when designing an algorithmic system. Table A.1 shows the layers of inequalities and what they mean. [46] maps these inequalities to prominent fairness philosophies discussed above, in a systematic manner.. We show this mapping in Table A.1.

### Considering a Socio-Technical Perspective

So far, we have an idea on the types of goals and objectives that can be outlined for fairness elicitation, that can be considered from literature. Certain scholars and researchers have criticized the neglect of the **social** considerations within Fair Machine Learning [63]. [63] performs a study to suggest constructive reforms to add this **social** perspective, into the technical implementations and processes of Fairness in ML. [63] investigating common traps that Fairness in Machine Learning research can fall into, whilst taking social notions into considerations. The authors follow up with a socio-technical perspective based solution to address these traps.

The socio-technical recommendation provided by [63] states that "when considering designing a new fair-ML solution, this would mean determining if a technical solution :



1. is appropriate to the situation in the first place, which requires a nuanced understanding of the relevant social context and its politics (Solutionism);
2. affects the social context in a predictable way such that the problem that the technology solves remains unchanged after its introduction (Ripple Effect);
3. can appropriately handle robust understandings of social requirements such as fairness, including the need for procedurality, contextuality, and contestability (Formalism);
4. has appropriately modeled the social and technical requirements of the actual context in which it will be deployed (Portability); and
5. is heterogeneously framed so as to include the data and social actors relevant to the localized question of fairness (Framing).

### Takeaways

Firstly, we see that fairness philosophies that can be understood to affect resource allocation, and affect algorithmic systems, can be used to define goals and objectives. [46] provides an overview of fairness philosophies that overlap and extend the fairness philosophies mentioned in [10]. We realize that the identification of these inequalities and bias can be treated as additional **goals and objectives** that can help map to a particular fairness philosophy (e.g. fair equality of opportunity, formal equality of opportunity, responsibility-sensitive egalitarianism).

We identify that many philosophies display forms of resource allocation strategies, similar to what distributive fairness entails (e.g. Luck egalitarianism, described above, states when redistribution of allocation is considered appropriate). This means that distributive and procedural fairness specifications and identifying fairness philosophies could yield in mentioning the same goals and objectives again. With this, we understand that multiple matrices can encapsulate social fairness goals and objectives, that are not necessarily mutually exclusive or derivative.

This means that multiple entities may be required to capture various types of goals and objectives, but it is not necessarily a mapping that leads to finding the correct social fairness specifications, rather an amalgamation of specifications that can lead to deeper thinking regarding social fairness notions.

For the framework, this means that one may chose to replace/remove/add/change social fairness notion entities as they are wish to, as long as it has justification on prompting social fairness specification for their systems. For example, an expert within sociology, or new research connecting sociology to algorithms may provide newer and improved social fairness notion entities.

We understand from the traps of *Solutionism*, *Ripple Effect* and *Portability* that a crucial aspect of eliciting fairness in the consideration of the social context that the technology is being deployed in. For the framework, this can mean that particular methods need to be included that help infer social context of the machine learning model. From the trap of *Framing*, the involvement of multiple social actors can be considered as well.

## 2.3 State-of-the-art in specifying fairness requirements

In this section, we want to identify fairness definition specifications that are empirical, and provide specificity. These fairness definitions can be inferred from the

Machine Learning community itself, due to their need to be empirical and implementable. As mentioned in [29], majority of the Machine Learning research in fairness is aimed towards Binary Classification. Hence, with the aim to understand how fairness is defined, categorized and measured in the ML community currently, we focus our analysis on fairness within towards resource allocation, specifically binary classification.

Firstly, binary classification tasks are relevant for our project as the subjects of these classification can be classified in particular outcome classes, that then result in resource allocation differences. For example, being the outcome of a loan-decision making model that affect the resources (in this case, the loan) that the subject to going to get. Secondly, limiting this study to binary classification is within the scope of this project, and does not hinder us from our aim to design a framework for F.R.E. This is because any study can be performed on a different high-level ML task, and the fairness definitions can be adapted to include any methods that allow for empirical fairness definitions to be specified.

To perform our search, we select four different survey papers/literature review studies, namely [45], [29], [6] and [77]. All these papers aim to categorize **fairness definitions and metrics in classification**, and provide an exhaustive view of this.

*Remark.* Interestingly, we note that each paper states there is a lack of clear guidelines on the **best** fairness definitions that can be applied to particular situations. To add to this, many fairness notions are impossible to apply together, making using fairness notions collectively quite challenging as well [29]. [77] also states the understanding the differences between these definitions itself can be quite difficult.

Continuing from inequalities and bias described above, the fairness definitions and metrics in ML are geared towards targeting these inequalities and removal of bias to ensure or mitigate fairness. These fairness definitions can be targeted to groups or individuals. We find that this target group for fairness can be used to specify whether the fairness definitions are group-based fairness, or individual fairness [50], [29]. Group-based fairness treat different groups equally, whilst individual fairness aims to give similar predictions to similar individuals.

Most of these approaches are based on the idea of **protected** or **sensitive** variables, and on (un)-privileged groups [29]. A protected or sensitive variable indicates a group in that should be treated fairly, For example, if we consider two groups, wherein the sensitive variable is gender, then fairness definitions and metrics are based on ensuring fairness between the protected group (where gender = female), and unprotected group (gender = male). One way of ensuring fairness for these groups is simply ignoring the sensitive variables during training. This is known as **Fairness through unawareness**. Technically, this can be considered a fairness definition. However, this approach struggles to tackle multiple sensitive variables, and **proxies**. Proxies are variables that can provide inference of the sensitive variables indirectly. For example, the sensitive variable of gender can have proxy variables of occupation, income or even working hours [].

With this basic understanding of what fairness, we can understand that, for example, we have two protected groups that are applying for a loan using the loan decision-making ML model. Let us say, that the protected variables here, are gender, and marital status. The protected group has gender = 1, and marital status = 1. The unprotected group has gender = 0, and marital status = 0. Hence, the protected group is disproportionately (less/more) likely to get positively classified. We know explore fairness definitions in the context of this example. [] performs a survey from papers stemming from NIPS, Big Data, AAI, FATML, ICML, and KDD, to create

	Actual - Positive	Actual - Negative
Predicted - Positive	<b>True Positive (TP)</b>	<b>False Positive (FP)</b>
	PPV = $\frac{TP}{TP+FP}$	FDR = $\frac{FP}{TP+FP}$
	TPR = $\frac{TP}{TP+FN}$	FPR = $\frac{FP}{FP+TN}$
Predicted - Negative	<b>False Negative (FN)</b>	<b>True Negative (TN)</b>
	FOR = $\frac{FN}{TN+FN}$	NPV = $\frac{TN}{TN+FN}$
	FNR = $\frac{FN}{TP+FN}$	TNR = $\frac{TN}{TN+FP}$

FIGURE 2.1: Confusion matrix of statistical notions [77]

an overview fairness definitions and metrics used within ML. [77] proceeds to categorize these definitions into statistical notions of fairness, similarity-based notions of fairness and casual reasoning.

### Statistical Notions of Fairness

Statistical notions of fairness mostly consist of definitions that are based on the confusion matrix generated by the classification. A confusion matrix [44] contains information about actual and predicted classifications done by a classification system. In Figure 2.1, we show the confusion matrix in relation to the protected and unprotected group. The particular terminology associated with it, is listed below :

1. **True positive (TP)**: a case when the predicted and actual classification are both in the positive class.
2. **False positive (FP)**: a case predicted to be in the positive class when the actual classification belongs to the negative class.
3. **False negative (FN)**: a case predicted to be in the negative class when the actual classification belongs to the positive class.
4. **True negative (TN)**: a case when the predicted and actual outcomes are both in the negative class.

All definitions under **statistical notions** utilize some form of comparison between these rates. When considering Figure 2.1, we also note that **Statistical Measures** can be categorized by whether they focus on **Predicted Classification**, **Actual Classification** and **Predicted Classification** or **Actual Classification** and **Predicted Probability**.

For example, statistical parity, categorized in **Predicted Classification** is satisfied if both protected and unprotected subjects have an equal probability of being assigned to the positive predicted class. Let us take a definition categorized under **Actual Classification and Predicted Classification**, for example Predictive Parity.

Predictive Parity is satisfied if both protected and unprotected groups have an equal positive predictive value. This represents the probability of the subject with a positive prediction to truly be in the positive class. In contrast to **Statistical Parity**, this takes into consideration both the predicted value and actual outcome. Another definition belonging to this category, that we can take a look at is **Predictive Equality**. Predictive Equality is satisfied if both the protected and unprotected subject have an equal False Positive Rate, which represents the probability of the subject in a truly negative class to gain a positive predictive value. Again, we can see how these definitions take into consideration the actual outcome and predicted value .

**Actual Classification and Predicted Probability** is similar, except it takes into consideration the predicted probability score rather than the value, in comparison to the outcome. For example, **test-fairness** is satisfied if for any predicted probability score  $S$ , the subjects have equal probability to truly belong to the positive class. And, **Well-calibration** builds on **Test-fairness** to state that "not both protected and unprotected groups should not only have an equal probability to truly belong to the positive class, but this probability should be equal to  $S$ ."

From [29], we find that these statistical measures can be segregated based on fairness criteria they satisfy, namely **Sufficiency**, **Independence** and **Separation**. While [77] separated the statistical notions based on the empirical calculation type, we find that here, the segregation is based on abstract fairness criteria.

Consider that  $S$  represents the sensitive variable, and  $R$  represents the classification score (between [0-1], for binary classification). Let  $Y$  represent the target variable. Then Independence represents non-discrimination independent of which group the subject belongs to. The limitation of independence is that it does take into consideration that  $Y$  may be correlated with  $S$ . Hence certain **fair** classifications may not be fair at all for a group. Separation tackles this limitation, by looks at whether the prediction score  $R$  and  $S$  are independent of the target variable  $Y$ .

### Similarity Based Measures

In statistical parity, as long as the protected and unprotected groups are assigned within probability in the positive predictive class, this is deemed fair. However, [77] motivates that if statistical parity is satisfied for female and male applicants, but female applicants receive this positive prediction due to them having high savings, but male applicants are selected at random, then this may still indicate unfairness.

Hence, while statistical measures consider the sensitive attributes alone, similarity measures address this by considering insensitive attributes as well. For example, **Casual Discrimination** is satisfied if the same prediction is made for two subjects with the same set of attributes. Or for example, **Fairness through awareness** is satisfied if two similar subjects produce similar classifications, where similarity is determined by a distance metric.

### Causal Reasoning

Lastly, definitions under casual reasoning pertain to capturing the relations between attributes and their influence on the outcome. These attributes can be referred to as **proxy** and **resolving** attributes wherein both types of attributes have relationships to the protected attribute, making the protected attributes **inferable** to the machine learning model. To tackle this, one definition is, **Counterfactual fairness**. **Counterfactual Fairness** states that "A causal graph is counterfactual fair if the predicted outcome  $d$  in the graph does not depend on a descendant of the protected attribute".

With this, we get a basic understanding on the different definitions and metrics within fairness in machine learning.

In terms of selecting the appropriate fairness definition, we find that two works, namely [2] and [61]. [2] provides a fairness tree that guides through the selection of certain fairness definitions. [61] provides a similar fairness flowchart for common fairness definitions of Unawareness, Individual Fairness, Statistical Parity, Equalized Odds, Disparate impact. Both, [2], and [61] provide a limited way of encapsulating the goals and objectives of fairness. For example, while certain aspects of distributive and procedural fairness specifications can be seen, the extensiveness of

goals and objective specifications identified in Social Fairness Notions (e.g. Table A.3, Table A.1).

Furthermore, [45] categorizes fairness definitions into the criteria that they address. With this, we can see the necessary criteria that may be important to define and identify for particular fairness definitions.

### Takeaways

We realize that fairness metrics (mathematical/empirical notations) and fairness definitions within ML can be considered entities. Fairness criteria (Sufficiency, Independence and Separation), target group (group-based/individual/counterfactual) and factors can be used to identify the relevant criteria that needs to be specified or known for achieving the fairness definitions and metrics. We do not consider these criteria to encapsulate fairness goals and objectives, but more like conditions of application for achieving particular fairness definitions and goals.

## 2.4 State-of-the-art in modeling fairness requirements

[11] builds and tests a fairness-aware machine learning pipeline, to show that satisfaction of all fairness metrics is impossible. This is also known as the impossibility theorem, which is proven in [62] which states that satisfying more than one of three fairness metrics is impossible, and requires the model developer to make trade-offs based on the context of the model, and which fairness metric is deemed more important to fulfill. Papers highlight the need for trade-offs to be made such as fairness-accuracy and so forth [15].

This means that the act of *specification* of fairness requirements may involve identifying and making decisions on particular trade-offs, which means that fairness requirements may need to be re-specified. We also note that these trade-offs can require building different models based on different fairness metrics to compare, as concluded in [11], indicating that this re-specification may in fact be an iterative process. Furthermore, when considering the nature of fairness requirements being uncertain and variable, trade-off management is expected to be iterative over time as well.

We find that [45], shown in A.3, proposes a specification on different fairness dimensions which can help in conflict resolution. We can leverage this A.3 to prompt dimensions on which trade-offs can be for-seen or recorded.

To be able to record trade-offs in these dimensions, we see that various stages of the model may incur a trade-off. For example, *internal data quality* and *model performance* dimensions within Table A.3 can induce trade-offs in the data-processing stage of the machine learning pipeline and the model evaluation stage of the machine learning pipeline. This prompts us to investigate on the stages within the machine learning pipeline, and how technical mitigation methods may be applied to each stage. With this information, we can understand the granularity of how trade-offs may be recorded. Looking beyond trade-offs, we can understand the granularity of the fairness steps that may be taken.

### Addressing the Machine Learning Pipeline

To identify the entities within the machine learning model, we wish to gain an overview of the types of models being built in ML, and the pipeline involved in

building them (i.e. the stages of ML). With this, we can understand how we can accommodate the ML model workflow with our framework.

### Stages in Machine Learning Model Pipeline

We traverse literature to identify the stages of development of a ML model. [4] lists a comprehensive outline of the stages within a ML workflow. They state that some stages are data-oriented, and others are model-oriented. Data-oriented stages include collection, cleaning, and labeling). Model-oriented stages include model requirements, feature engineering, training, evaluation, deployment, and monitoring. [49] specifies the following stages within Machine Learning :

1. **Model requirements stage** which is related to the agreement between stakeholders and the way the model should work.
2. **Data processing stage** which involves data collection, cleaning and labeling (in case of supervised learning).
3. **Feature engineering stage** which involves the modification of the selected data.
4. **Model training stage** which is related to the way the selected model is trained and tuned on the (labeled) data.
5. **Model evaluation stage** which regards to the measurements used in order to evaluate the model.
6. **Model deployment stage** which includes deploying, monitoring and maintaining the model.

[24] defines the workflow of the ML in a similar way of categorization, data handling (pertaining to data-oriented, model-oriented from [4], and data-processing and the three model stages from [49]). We do note that within the **data handling** categorization, [24] goes on to list Data Acquisition, Data Labeling, Data Exploration, Data Structuring and Feature Engineering. This lists more stages than was mentioned by [4], for data oriented tasks.

From this, we see that the stages within the ML pipeline remain more-or-less similar. While some may define stages with slightly higher granularity, the stages still pertain to being either data oriented, model oriented. There we can treat the stages of ML as entities, and understand that these stages may even be pre-defined.

One additional interesting insight that comes to light is that these workflows are not strictly sequential, but rather iterative. These iterations occur after the evaluation step and can result in iterative execution of the model and the data-oriented stages [4], [24]. This lets us know that enforcing sequentially in the pipeline may not be so relevant for ML models.

### Activities of Machine Learning Models

Machine Learning can be seen as an intersection between statistics and computer science. It is based on this idea of **training**, extracting information from data-sets by finding underlying patterns using various statistical techniques [7].

While training can be seen as a primary target, achieving it heavily depends on the data and desired outcomes. These activities can vary in terms of the type

of learning style used within training. These learning styles can be categorized as follows [7]:

1. Unsupervised Learning
2. Supervised Learning, semi-supervised learning
3. Re-enforcement Learning

With the popularity of ML, we see a few popular tasks that are associated with each of these learning styles.

**Unsupervised learning** can involve tasks such as clustering and prediction. A few common algorithms facilitating this learning style involve K-means, Gaussian Mixture Models and Dirichlet process mixture mode. [56]

**Supervised learning** can involve tasks such as classification, regression and estimation. A few common algorithms facilitating this learning style involve Naive Bayes, Support Vector Machines, Bayesian Networks and Neural Networks. [56]

**Re-enforcement learning** can involve tasks such as decision making. A few common algorithms facilitating this learning style involve Q-Learning, R-Learning and Sarsa Learning.

We also see a few advanced learning styles appear to manage the vastness and complex nature of big data. This learning styles include representational learning, Distributed and Parallel Learning and Deep Learning. [56]

**Representational Learning** is designed to improve efficiency when to aid with high-dimensionality in data, both in terms of results and computational time. Representational Learning aims to optimize the input configurations in captures with the learned representation. Feature selection, Feature Extraction and Distance Metric Learning are an effort to perform representational learning.

**Deep Learning** allows for deep architectures to often infer high-dimensional, complex, hierarchical patterns within data. Due to the **deep** nature of these techniques, they can outperform traditional representation selections, and hand-made solution to provide adaptive and efficient solutions. Deep Learning has gained traction due to its role in Computer Vision, Natural Language Processing and Information Retrieval. Deep Learning nods to Evolutionary AI as well, a research field that is slowly gaining popularity. [56]

**Distributed and parallel learning** allows for big data to be learned **simultaneous/iteratively** in an efficient manner. Most learning methods are constricted to being able to utilize all data at once and learn on them. Taking a distributed approach to the learning aims to solve this issue and allow for learning on the maximum of the data. Decision rules, staked generationalization and meta-learning are techniques of this type of learning. [56]

**Transfer learning** allows for learned information extracted during a certain task to be applied to different tasks. This means that previously learned information can be used, hence improving efficiency of the task. [56]

### Take-aways

We identify that there can be various combinations of tasks, types of learning and models used within Machine Learning Models. For example, one can perform the task of **binary classification**, using neural networks or a simple logistic regression model. We can see the the activities within model evaluation, will differ based on

Machine Learning Stages	(Example) Machine Learning Activities
Data Processing	Compression, Data Encoding, Data Collection
Feature Engineering	Correlations, Feature Encoding, Feature Selection
Model Training	Baseline Classifier, Optimization
Model Evaluating	Confusion Matrix, AUC Score
Model Deployment	A/B Testing

TABLE 2.1: Machine Learning Model - Stages and Example Activities

the model. We also see that, based on the type of learning, **unsupervised** or **supervised**, the processing of the data can differ. Furthermore, with the advanced learning styles, such as **Representation Learning** and **Deep Learning**, the feature engineering activities can differ. For example, in representation learning, feature selection and feature extraction are explicit activities, whereas in deep learning, these feature selection and extraction is embedding within the model itself due to its nature of being able to address high dimensionality of the data.

In Table 2.1, we can see an example of different activities that might be entails within different types of models.

Combining the insights on the types of machine learning models, we find that while the stages within the workflow of Machine Learning models are more or less standard, the activities conducted within these stages conducted for the models can be dependent on various factors such as data sources and objectives of the model.

Going back to the questions we defined at the start of this section, namely "What entities are there within the ML model", "How do these entities interact and amongst themselves", we find that both stages and activities can be defining entities in ML models, and activities can be classified into a particular stage(s).

## Technical Mitigation Methods for Fairness in Machine Learning

In this section, we explore technical mitigation methods of fairness within ML. This will give us an indication of how the fairness entities and ML entities, uncovered in the previous section, can interact. While, fairness metrics are empirical formulations of implementing various fairness definitions, our goal is to understand techniques within fairness that facilitate these fairness metrics and to look beyond these metrics as well.

To this end, we study survey literature targeted towards technical mitigation methods of fairness within ML (specifically, binary classification). [10] state that most technical mitigation methods can be applied to data preparation, model training or post-processing stages.

[29] studies over 300 papers to provide a comprehensive study of the technical mitigation techniques within the pre-processing (i.e. data preparation), in-processing (i.e. model training) and post-processing stages. In Table A.4, we show the mitigation techniques, brief description and the stages categorized in [29].

## Conclusion

From the background study, we create an overview of the state-of-the-art practices we find in fairness requirements elicitation, specification and modeling. This is shown in Figure 2.2.



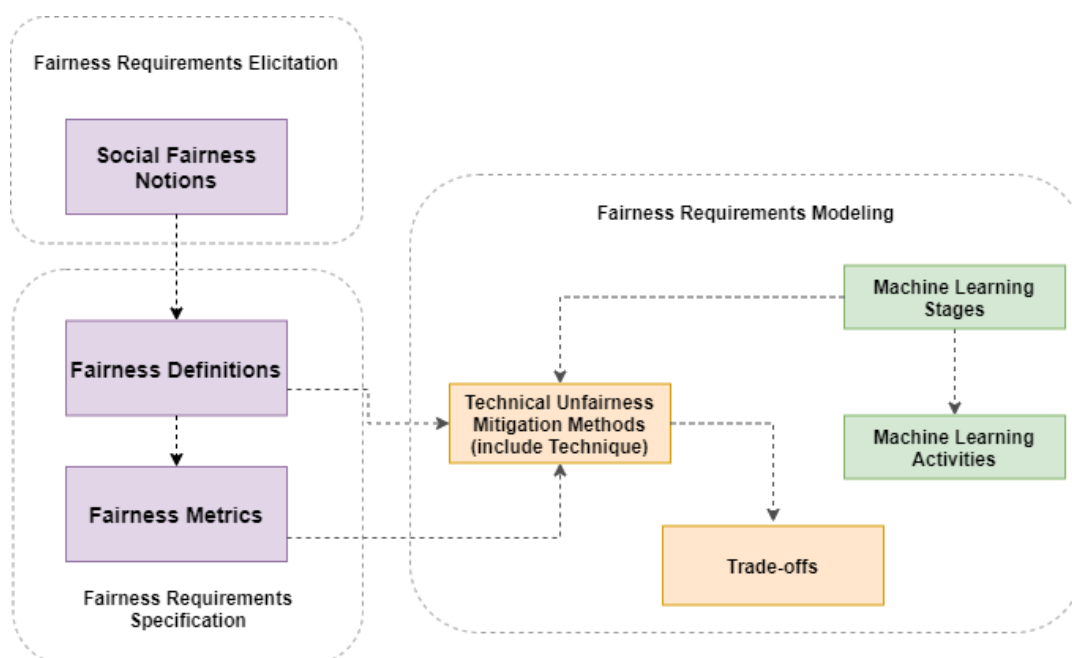


FIGURE 2.2: High Level Overview of Findings from Background Study



## Chapter 3

# Framework Requirements Elicitation and Specification

In this Chapter, we tackle Research Question 2, and 3a. We want to gain in-depth analysis of what the status quo of fairness requirements engineering looks within industry (in terms of understanding how fairness goals, objectives and motives are communicated), and what the desired fairness goals, objectives and motives for an ML model may look like. We want to reflect on this, and derive requirements to design the framework. We traversing the objectives, devise a methodology and obtain results for this chapter, for specifically:

### 1. Objectives:

- (a) **Challenges** : We infer institutional challenges within F.R.E.M. by understanding the following :
  - i. **The level of Awareness** - The level of knowledge regarding fairness goals, objectives and motives
  - ii. **The level of Transparency** - The communication procedures regarding fairness goals, objectives and motives
- (b) **Requirements** : We derive requirements for the M.L.F.F. by reflecting on the notations required for performing requirements elicitation, modeling and reflecting on the institutional challenges (and findings) discovered in the previous objective. *A notation can be regarding as a particular outcome to be noted when performing requirements elicitation and modeling.*
- (c) **Insights** : We derive insights to consider for the M.L.F.F. by reflecting on the best methods, strategies and techniques for requirements elicitation, modeling and reflecting on the institutional challenges (and findings) discovered in the previous objective.

### 2. Methodology:

- (a) **Challenges** : We utilize the institutional context, and perform stakeholder studies, wherein the stakeholders belong to the ML model development process within the institution.
- (b) **Requirements** : We utilize [17], that performs an exhaustive survey study on *Notations* for requirements elicitation and modeling traverse through the relevant literature cited, and reflect on the relevancy, institutional challenges and derive requirements.
- (c) **Insights** We utilize [17], that performs an exhaustive survey study on *Methodology, Strategy and Advise* for requirements elicitation and modeling

traverse through the relevant literature cited, and reflect on the relevancy, institutional challenges and derive insights.

### 3. Results:

#### (a) Challenges

- i. **Status Quo procedures regarding F.R.E.M.** : that are analyzed via the stakeholder studies at ING.
- ii. **Challenges related to F.R.E.M.** : that are analyzed via the stakeholder studies at ING.
- iii. **Dimensions to address for F.R.E.M.** : from the challenges, we can infer the dimensions we need to address. We keep these dimensions as general as possible, to avoid being too institution specific.

### 4. Requirements

#### (a) List of requirements for the framework

### 5. Insights

#### (a) List of insights for the framework

## 3.1 Institutional Challenges

To recognize the challenges within the institution regarding fairness requirements engineering, we perform stakeholder studies. We start off by identifying the stakeholder roles that we infer from our institution. We then identify objectives that we would like to infer from these stakeholder roles, in regards to their knowledge and procedure (status quo, gaps and challenges) in terms of fairness requirements engineering in ML models.

We look towards identifying the baseline stakeholders for ML models being developed in the institution. We investigated that this particular institution has three categories of roles that baseline stakeholders could fall into :

#### 1. Model Owner

A model owner is the one in charge of requesting the model and ensuring its delivery. They also hold authority in making key decisions regarding the model.

#### 2. Model Developer

A model developer is working on the technical part of the model. This can also be any aspect related to the model (i.e. data, evaluation methods), at any point within the issuing of the model, and the delivery of the model.

#### 3. Domain Expert

This person can have in-depth knowledge of the application domain that the model will be deployed in. They can also be the people that will be using the model.

We construct an overview of these roles in Figure 3.1. Our goal is to design objectives for the study in regards to the role's knowledge and processes, and eventually populate the Figure 3.1 with the insights of those objectives, gained from the stakeholder studies. We focus the objectives of the study, namely **Level of Awareness** and **Level of Transparency**. Regarding **Level of Awareness**, we would like to understand the current knowledge that exists regarding fairness in the roles. In terms of **Level of Transparency**, we want to build further understanding on the role's processes and interactions, communication objectives and struggles (in relation to fairness in ML).

We break-down the two key objectives, Level of Awareness and Level of Transparency, into sub-objectives to motivate further.

### Level of Awareness

1. **Objective 1** What current knowledge (including its quality and its application) is held by the stakeholders in regards to fairness?
  - (a) **Objective 1a** How extensive is their knowledge on where fairness issues might occur?  
This will tell us about what assumptions the tool can make regarding the base knowledge present within the sector
  - (b) **Objective 1b** Can the stakeholders identify the not-so-obvious 'red flags' that can occur regarding fairness in ML?  
This will give us insight on to what extent and granularity the fairness is accessed

### Level of Transparency

1. **Objective 2** *What current struggles or gaps are in between our stakeholders' communication in regards to fairness?*
  - (a) **Objective 2a** How are they going to communicate that something is fair? And, when fairness can mean multiple things, what does this look like? How formal do they need this procedure to be? This will tell us how we should structure fairness notions within the tool. As we are dealing with social context, formalization can result in removing subjectivity from the information. By definition, social context is contextual so the challenge is to facilitate navigating this line of subjectivity and formalization
  - (b) **Objective 2b** What do they each view as accountable/responsibility? This will tell us how the tool needs to distribute knowledge amongst the stakeholders
  - (c) **Objective 2c** What are they currently doing to show people that societal expectations are being met? Is it a defensive or an aggressive approach. Here, we try to get more information on how we can bridge this link between automation and human experts.

We expect that multiple of these objectives to infer other aspects. For example, combining Objective 1(a) and 2(b) can give us an indication regarding the extent to which each stakeholder is willing to go in order to ensure fairness. Similarly, the relationship between 1(a) and 1(c) can let us know about some 'hidden' gaps that still need to be tackled within the communication.

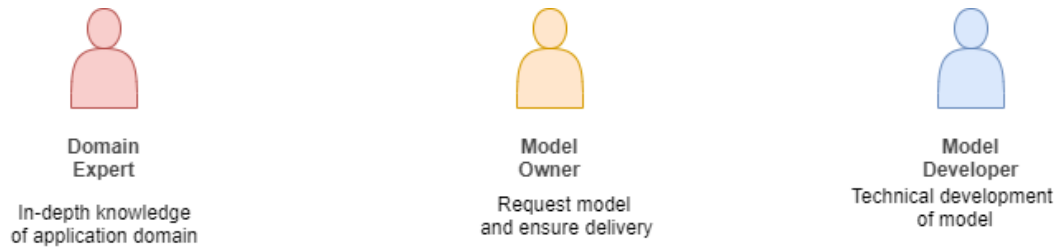


FIGURE 3.1: Roles within Stakeholder Study for ML Model Development

### 3.1.1 Stakeholder Study

Now, we move towards designing our stakeholder study. To gain rich and correct information through these stakeholder studies, we need to ensure that our design is well thought out. To accomplish this, we review a few papers to compare different techniques used for requirements elicitation involving contact with users (i.e. our stakeholders). We, again, use the approach of finding these papers via a few survey papers, to ensure we take a more exhaustive and systematic approach.

#### Techniques for conducting Stakeholder Study

[58] provides an overview of techniques for *trawling* requirements, based on interactions with stakeholders. *Trawling* means fishing or netting. [58] uses this terminology to elude to the *fishing* of requirements. This *fishing* is due to the nature of the requirements. [58] introduces three natures, namely:

1. **Conscious**

"A conscious requirement is something the stakeholder is particularly aware of."

2. **Unconscious**

"When a stakeholder does not mention a requirement because he does not realize that he has it (i.e. the requirement appears so trivial to the stakeholder, that there is no communication on it)."

3. **Undreamed**

"Requirements that do not even occur to the stakeholders because they cannot imagine what it"

From this, we understand that directly asking requirements to the stakeholder may not capture all relevant requirements. We look towards techniques that can facilitate the inference of these three natured requirements. In [66], the author provides us with an overview of how different techniques may bring-out surface level or deeper level inferences from stakeholders. This overview is shown in Figure 3.2.

As we see from Figure 3.2, the levels *say, think* relates to *Conscious*, *do, use* relates to *unconscious* and *know, feel and dream* related to *undreamed*.

Keeping this in mind, we explore various methods that can accomplish these three natures. We display the promising techniques we find, and discuss them. By promising, we mean it is suited to the project, aligned with one of the three natures and whether it is feasible to conduct within the participatory institution.

The promising techniques includes :

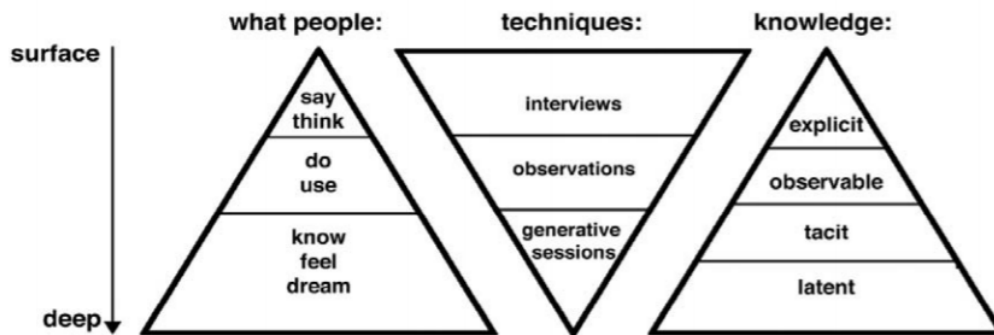


FIGURE 3.2: Overview of different levels of knowledge, accessed by different methods [66]

### 1. Interviews

Interviews are the most common way of inferring direct requirements from stakeholders. This can help identify *conscious* requirements. 3.2

### 2. Soft Systems

[16] introduces a framework of modeling soft systems. This framework consists of :

- (a) Customer - The beneficiaries or victims of the system.
- (b) Actors - those who carry out the transformations within the system.
- (c) Transformation - of some defined input to defined output.
- (d) Weltanschauung - the image of the world that makes this system meaningful.
- (e) Owned - the owner/s of the system.
- (f) Environment - the environmental constraints.

Soft systems inquiry can allow for a broader world view to be established, and identify requirement needs early on in the process. This can help refer to all *conscious* and *unconscious* requirements.

### 3. Simulation Models

A simulation model refers to a model where conditions are mimicked from a real-life case. [69] argues that if these simulations are based on the stakeholders' knowledge and perception, then one can acquire requirements that may be overlooked until the development of the product has been done. One way to achieve this, is by including details in the simulations that pertain to stakeholders' characteristics of their role.

For our project, these simulations are going to be paper-based, as the simulations include interactions within the institution and with their end-users, rather than any specific environment. This can help refer to all *conscious* and *unconscious* requirements.

### 4. Viewpoints

[60] speaks on acquiring viewpoints, namely the viewpoint of the current world, the viewpoint of the future world of the product being engineered for. This allows for a more focused and systematic approach to inferring requirements. This can help refer to all *conscious*, *unconscious* requirements and *undreamed* requirements

## Execution of Stakeholder Studies

### Experimental Setup

We gain a lot of inspiration from the research conducted above, where we display the promising techniques to be used within the design of the study.

We also make other considerations regarding the design. These considerations stem from an awareness that the topic of this project, namely *Fairness in Machine Learning*, can be a highly sensitive topic, specifically within an existing real-life institution. We also consider practical considerations such as time, and quantity of participants.

We want to avoid getting defensive answers and focus on questions that can help identify struggles that our stakeholders are facing. Lastly, all interviews consist of a time constraint hence we estimate that we will get an opportunity to pose five to six questions to each stakeholder.

The quantity of participants within the study will be limited. This means that quantifiable analysis may not be possible or significant. Therefore, qualitative analysis may have to be conducted. Hence, ensuring information richness becomes key here. So the design needs to allow for probing the stakeholders' answers in order to ask follow-up questions.

For the final stakeholder study, we select a **semi-structured interview** with **aimed duration of thirty minutes** which is based on the stakeholder traversing through a *case*, which we will refer to as a case study. This *case* is designed around if *unfairness* were to occur using their models.

Each case study is customized to pertain to the domain of application and stakeholder background. The content of the stakeholder studies can be referred to in Appendix B. In Table B.1, we provide the setup of the stakeholder studies with a mapping to the relevant Appendix section, number of stakeholders per role and the high-level ML task of the model. In Table 3.2, we show the number of questions in each stakeholder study that investigate the objectives described in Section 3.1. For an elaborate mapping that shows which objectives each question accomplishes, one can refer to Table B.1 in Appendix B.

## Results

### Stakeholder Study Analysis

To analyze the interviews, we reflect on the objectives defined in Section 3.1. We use Objective 1 and 2 (*Current Knowledge* and **Stakeholder Communication**).

We perform this analysis on a per role basis. This allows us to infer information across the same role (i.e. Model Developer, Domain Expert or Model Owner) across models. We also perform analysis on a per model basis. This allows us to infer information on a particular model basis. We can then cross-tally this with other models' analysis to see if we find any over-arching observations as well.



Stakeholder Study Roles	Model	Task of Model
Stakeholder Study A (Appendix B)	Binary Classification	1 Model Owner, 2 Model Developers, 2 Domain Expert
Stakeholder Study B (Appendix B)	Entity Detection	1 Model Owner, 1 Model Developers, 1 Domain Expert
Stakeholder Study C (Appendix B)	Natural Language Processing	1 Model Owner, 1 Model Developers, 1 Domain Expert

TABLE 3.1: Stakeholder Study Setup

Task of Model	Roles	Objective	Number of Questions
Binary Classification	Model Owner	Objective 1a	2
Binary Classification	Model Owner	Objective 1b	2
Binary Classification	Model Owner	Objective 2a	1
Binary Classification	Model Owner	Objective 2b	1
Binary Classification	Model Owner	Objective 2c	1
Binary Classification	Model Developer	Objective 2a	2
Binary Classification	Model Developer	Objective 1a	4
Binary Classification	Model Developer	Objective 1b	4
Binary Classification	Model Developer	Objective 2b	2
Binary Classification	Model Developer	Objective 2c	1
Binary Classification	Domain Expert	Objective 1a	2
Binary Classification	Domain Expert	Objective 1b	4
Binary Classification	Domain Expert	Objective 2a	2
Binary Classification	Domain Expert	Objective 2b	1
Binary Classification	Domain Expert	Objective 2c	2
NLP, Entity Detection	Model Owner	Objective 1a	3
NLP, Entity Detection	Model Owner	Objective 1b	4
NLP, Entity Detection	Model Owner	Objective 2a	1
NLP, Entity Detection	Model Owner	Objective 2b	1
NLP, Entity Detection	Model Owner	Objective 2c	2
NLP, Entity Detection	Model Developer	Objective 1a	4
NLP, Entity Detection	Model Developer	Objective 1b	3
NLP, Entity Detection	Model Developer	Objective 2a	2
NLP, Entity Detection	Model Developer	Objective 2b	2
NLP, Entity Detection	Model Developer	Objective 2c	1
NLP, Entity Detection	Domain Expert	Objective 1a	2
NLP, Entity Detection	Domain Expert	Objective 1b	3
NLP, Entity Detection	Domain Expert	Objective 2a	1
NLP, Entity Detection	Domain Expert	Objective 2b	1
NLP, Entity Detection	Domain Expert	Objective 2c	1

TABLE 3.2: Summary of Stakeholder Study Mapping to Objectives

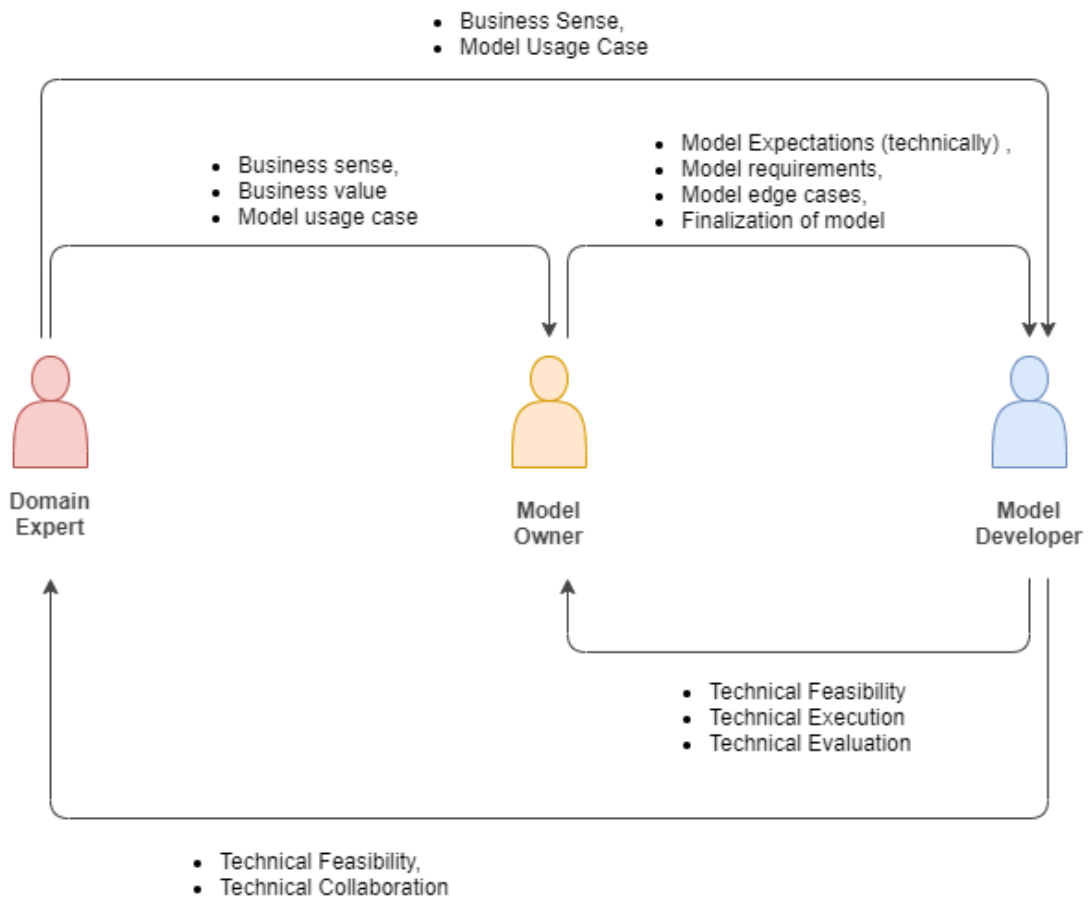


FIGURE 3.3: Status Quo Procedure for Communication between stakeholders

The complete list of codes and statistical data are displayed in the Appendix. Below, we present the interesting codes that we find, and how we see them emerging. We will go through this, per objective.

*Remark.* The sub-objectives defined in Section 3.1 motivate the creation of the codes. However, we found that defining codes that combines the sub-objectives results in more meaningful insights to understand the procedure as a whole. Hence, there is no clear categorization of codes for the sub-objectives, but only for the two objectives themselves.

Due to confidentiality purposes, we refrain from displaying any transcripts of the interviews. We only display the results of the coding statistics.

### Status Quo Procedures of F.R.E.M.

As described at the start of this section, we started off with Figure 3.1, and with our stakeholder studies, are able to populate it, shown in Figure 3.3, and Figure 3.4. Let us walk through the procedures discovered through our stakeholder studies. This will be applicable in understanding the requirements of the framework, as we can design to fit the general procedures (and gaps within) of the stakeholders.

In Figure 3.3, we present the visualization of the status quo stakeholder communication properties for building an ML model. In Figure 3.4, we show the status quo communication and responsibilities assumed regarding fairness.

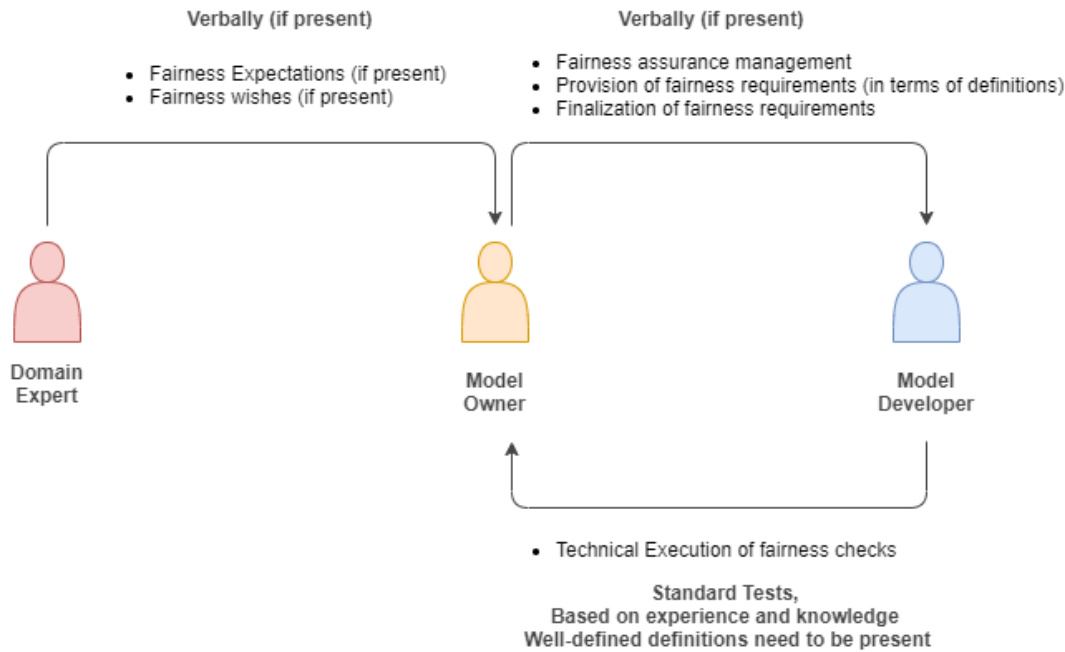


FIGURE 3.4: Status Quo Procedure for Fairness Responsibilities between stakeholders

In Figure 3.3, we find that the domain expert conveys the business sense behind data and features, to the Model Developer. The domain expert provides this to the model owner as well, in addition to the business value. Here, we infer that the discussion's regarding the model's outcome and accomplishments occur in this part of the communication. The Model Owner is then in charge of communication the decisions on what the model should outcome, and the technical capabilities of the model to the Model Developer. This can be done in the form of specifying requirements. The Model Owner has final say on decisions regarding the model and the technical expectations, for example what accuracy may be acceptable, what the time constraints for building the model are and so on. The Model Developer is then in charge of the technical execution and evaluation of the model, pertaining to the direction provided by the Model Owner. The Model Developer can have conversations regarding technical feasibility and so forth with the Model Owner. Model Developers can further communicate to the domain expert regarding getting business sense out of data, or perhaps certain technical feasibility.

Narrowing down this communication to Fairness specifically, Figure 3.4 shows the fairness communication. The domain expert can extract relevant fairness issues from their experience and communicate it to the model owner. The model owner then has responsibility on deciding the relevant fairness issues, and what the model should implement. We find that they have to consider both societal expectations and the institutional expectations. The model developer has responsibility on ensuring the fairness is implemented technically.

We find that most of the communication regarding fairness is verbal. Further we find that there may be misconceptions, limited knowledge and lack of formalization on fairness definitions and technical mitigation of fairness. We also find that model developers require a concrete explanation of what fairness is, from the model owners. This is because, they can find that implementing fairness is such a broad term that it can be quite impossible to navigate through.

With this, we understand the current level of awareness and transparency within

the procedures regarding fairness. Moreover, we deduce the need for added awareness and transparency within the procedure to ensure fairness considerations.

Hence, we go through our two main objectives, Level of Awareness and Level of Transparency, and list observations pertaining to each objective, and derive requirements from these observations. The approach for listing these observations takes into consideration the inferences we make above, that gear towards improving the Level of Awareness, and Level of Transparency whilst abiding to the general backbone of the procedure and interaction of roles discovered.

### **Institutional Challenges to address in the framework**

For Objective 1 and Objective 2 mentioned in Section 3.1, we list observations from the stakeholder study analysis, along with what these observations indicate regarding conducting F.R.E.M.

### **Objective 1 : Level of Awareness**

1. **Observation A.1** Social fairness concepts and dilemmas can trigger discussion.

Interestingly, we discover that participants may have varying knowledge regarding fairness, and more importantly conflicting ideas of what fairness should look like within the model. One question we use to test this in our case study, is being translating the confusion matrix into lei-man terms for the domain expert and model owner, and then inferring their perspective.

For example, for the loan decision model, we ask :

- (a) Of those to whom I granted a loan, how many will actually not pay?
- (b) Of those that I decided to reject, how many would actually pay?

We find that by simply displaying various fairness concepts, and asking people from different backgrounds, involved within the construction of the ML model, can bring them to think deeper into what actually would be *fair*.

**Facilitation of F.R.E.M.:** This helps in elicitation as they provoke deeper consideration on devising of fairness goals, objectives and motives

2. **Observation A.2** Limitations and opportunities for fairness definitions can be communicated.

We find that technical feasibility needs to be communicated regarding fairness definitions. We know that there is a lack of knowledge regarding technical mitigation methods. From our background study in Chapter 2, we discover that there can be many trade-offs regarding these technical mitigation methods.

**Facilitation of F.R.E.M. :** This helps in elicitation as they provoke deeper consideration on the trade-offs of fairness goals, objectives and motives.

3. **Observation A.3** Fairness misconceptions can be communicated.

For example, people may consider *Fairness through unawareness* to be a valid solution for fairness, not realizing the impact of proxies.

Regarding ML models themselves, we find that the participants can fail to see the causes of unfairness in these models, or reflect on situations where unfairness could occur.

From our background study, we see that there exist conditions of application and limitations to technical mitigation methods. In a few papers, we find the authors prompting the developers on which is the correct mitigation method for their purpose.

**Facilitation of F.R.E.M. :** This helps in elicitation as they provoke deeper consideration on devising of fairness goals, objectives and motives.

4. **Observation A.4** Mitigation methods can be highlighted.

We find that the knowledge on fairness mitigation methods pertaining can vary depending on the developer. We also find that developers are aware of standard statistical measures, and may apply those for ensuring fairness.

From our background study, we find that the mitigation methods may defer based on the ML pipeline stages and the different activities conducted. From the stakeholder study, we find that the activities and stages can be distributed through different developers, and each developer may want to infer to only the mitigation methods that pertain to their particular activity or domain.

**Facilitation of F.R.E.M. :** This helps in modeling as conditions of application can be seen with the mitigation methods, to identify the constraints of fairness goals, objectives and motives.

5. **Observation A.5** Formalization of fairness definitions can be communicated.

From our stakeholder studies, we find that seeing the nuances in fairness definitions can be difficult for the stakeholders.

**Facilitation of F.R.E.M. :** This helps in modeling as relationships between fairness definitions to identify the constraints of fairness goals, objectives and motives.

## Objective 2 : Level of Transparency

1. **Observation T.1** Stakeholders need to effectively communicate fairness requirements in terms of their own background and knowledge level.

We find that the specific roles in our study, have specific backgrounds and knowledge levels. We want to facilitate transparency by each role communicating their specific responsibilities within the framework. This can be used to infer the responsibilities and the gaps that may occur. We find that within the interviews, since the business sense and amalgamation of context for fairness can come from the model owner and domain expert.

**Facilitation of F.R.E.M. :** This helps in elicitation as it allows for familiar terminology to be used when devising of fairness goals, objectives and motives.

2. **Observation T.2** Institutionally, communication of decisions made or considered should be recorded for each model.

We find that consideration that fairness might be occurring at different stages of a pipeline, involve cross-models and different people within the institution can be considered. For example, We infer that stakeholders are primarily considered about their own model. If other models are being used to build a model, then the reliance of information is mostly on the documentation attached to these other models.

**Facilitation of F.R.E.M. :** This helps in elicitation and modeling as relationships, behaviors and entities of fairness goals, objectives and motives can be seen through multiple models and pipeline stages.

3. **Observation T.3** Stakeholder beliefs can be captured to infer their perspectives and viewpoints.

We find that as stakeholders are exposed to different levels of information, their perspectives may change. From our background study, we also figure that decisions can change over time, or be influenced by other factors.

**Facilitation of F.R.E.M. :** This helps in elicitation as it provoke deeper consideration on devising of fairness goals, objectives and motives

From the observations above, we see that Observations A.1, A.3 and T.3 show that by invoking **deeper consideration** of various fairness aspects, facilitation of F.R.E.M. can be improved. Observations A.2, A.4, A.5 and T.2 elude to the **need for specification** in order to perform F.R.E.M. Lastly, Observations T.1, T.2 and T.3 indicate that **prompting communication** among stakeholders by allowing the usage of familiar terminologies and constraining the communication itself (e.g. by filling out a pre-defined matrix) can help improve F.R.E.M. From this, we can identify three high-level challenges within the institution, namely :

1. **Consideration** : Prompting and thinking over fairness in ML models from different angles, involving social context and technical methods.
2. **Specification** : Specifying fairness in ML model in such a way that granularity is introduced in devising a fairness definition, to facilitate empirical definitions of fairness.
3. **Communication** : Communicating concepts and decisions within fairness requirements engineering can be difficult if common terminology is not present, and in accordance to the background of the stakeholders.

## 3.2 Requirements for the framework

In this section, we derive requirements for the framework. We analyze the best practices (notations, methodologies, strategies and advice) within requirements elicitation and modeling, while reflecting upon the institutional challenges and observations discovered in Chapter 3 to formulate requirements and insights to consider when designing the framework.

We will use [17] overview of the notations, methodologies, strategies and advise for requirements elicitation and modeling. This is displayed in Table 3.3 and Table 3.5. We choose [17], due to it being a survey-based paper, allowing us to perform a structured and somewhat exhaustive search. In Table 3.3, *Notations and Methodologies, Strategy and Advice* are high-level decomposition of solution-categories for requirements engineering stages. Furthermore, we reflect on methodologies, strategies and advise elude to possible solutions for implementing these notations. We extract interesting insights that can be relevant to consider when designing the framework.

F.R.E.M Stages	Notations
Requirements Elicitation	Goals, Policies, Scenarios, Agents, Anti-models,
Requirements Modeling	Object models, Behavioral models, Domain descriptions, Property languages, Notation Semantics

TABLE 3.3: F.R.E.M Elicitation, and Modeling Notations noted from [17]

## Notations

We traverse through the notations specified in Table 3.3, for each F.R.E.M Stage. We interpret notations as ways to encode information in a structured and meaningful manner. We align these notations with the fairness observations, and then specify behavioral requirements, should they be applicable.

## F.R.E.M Elicitation

### 1. Goals

The framework should facilitate the identification of goals, in terms of fairness. From **Observation A.5**, we know that these goals can be presented as different notions of fairness. From **Observation T.3**, we know that different stakeholders might have different goals and this needs to be communicated as well. From , we know that goals can change or be appended over time.

[47] speaks on goal-orientation being based upon agents. It states that the assignment of responsibilities for goals to agents is a critical decision in the requirements engineering process. Furthermore, it states that goals can be refined such that agents are assigned goals that are realizable. From **Observation A.2**, we see that limitations and opportunities can influence the existing goals. From **Observation A.3**, we see that misconceptions may dis-orient the view on limitations and opportunities, which in turn can lead to increasing the number of iterations required in re-consideration of the goals.

- (a) **M.1** Each stakeholder should be able to define their goals in formalized fairness definitions
- (b) **M.2** Each stakeholder should be able to get an understanding of how realizable the goals are
- (c) **M.3** Each stakeholder should be able to re-define their goals

## Policies

The framework should facilitate policies, in terms of fairness. From **Observation T.1**, we know that each model might have a different level of fairness in needs to adhere to. This level can be intertwined with institutional policies on the model implications. From **Observation T.2**, previous policies and decisions can be referred to and may reflect differences on policies and institutional

change of policies that occur through the organization. For example, sensitive data usage is not allowed within models in productions, but is permitted during the validation of models. Then the prevention of sensitive data usage can be a policy applicable to models within the data processing pipeline stage, but not relevant within the validation pipeline stage.

- (a) **M.4** Each stakeholder can view policies on an institutional level
- (b) **M.5** Each stakeholder can view policies on an model level
- (c) **M.6** Each stakeholder can reflect on the model level policies
- (d) **M.7** Each stakeholder can discuss the model level policies

## 2. Agents

The framework should allow different agents to be consulted when creating the requirements. From TR5, agents can be stakeholders with different backgrounds and authority within an institution. From **Observation A.1, A.3 and T.3**, there should be active facilitation of discussion and reflection between these agents.

- (a) **M.8** Stakeholders with different backgrounds need to be able to use the framework
- (b) **M.9** Stakeholders with different backgrounds need to be able to view each other's viewpoints

## 3. Anti-models

Anti-models can be used to detect vulnerabilities and potential threats during the requirements engineering phase itself. One example of an anti-model is described in [75]. It is developed on the basis of Message Sequence Charts (MSC). MSC is a popular way of requirements elicitation and specification. In MSC, the system architecture and intended system behavior is outlined. When there is a mis-match between the system's intended behavior and architecture, then this gives rise to an **Implied Scenario**. Implied scenarios occur because a component's local view of the system state is insufficient to enforce specified system behavior. [73] go in-depth about detecting these implied scenarios, and [75] provides a system for encapsulating the negative scenarios that may occur.

From **Observation A.1** and **Observation T.1**, we understand that the encapsulation of situations and scenarios can be helpful to refer to when defining requirements. It can also facilitate AR6, the addition of new information, methods and techniques in order to deal with newly uncovered situations.

- (a) **M.10** Stakeholders need to be able to review scenarios when certain model behaviors occur
- (b) **M.11** Stakeholders need to be able to add scenarios when certain model behaviors occur

## 4. Non-functional Requirements

Non-functional requirements (NFRs) are describe the constraints on the solution space, and capture a broad spectrum of properties such as reliability, portability, maintainability, usability, safety, and security [78]. [20] uses the idea of formulating NFRs as soft-goals and defining some form of traceability



to goals. For many institutions, fairness itself can be considered a NFR, specially as it stems beyond an algorithmic adjustment. While **Observation A.5** focuses on the formalization of fairness requirements to make them as functional as possible, we can also consider facilitating some form of NFRs within the framework. We can see scenarios and discussion **Observation A.1** and *Observation T.1* can promote the inference of NFRs. Due to the goal of this framework being around making fairness requirements as functional as possible, we scope the NFRs to inference only. The implications of this can be checked in the evaluation of the framework.

- (a) **M.12** Stakeholders need to be able to infer NFRs from the framework

## F.R.E.M. Modeling

### 1. Object models

[27] defines object modeling notation as a simple kind of first-order specification language. Object models describe state spaces in which the individual states are structured with sets and relations. They form the backbone of most object-oriented development approaches, and of efforts in model-driven architecture. The problem that **Model-driven architecture** tries to address by stating 'Most IT organizations rely on a complicated assortment of infrastructure technologies that have evolved over multiple years. To develop applications in this context requires an approach to software architecture that helps architects evolve their solutions in flexible ways, reusing existing efforts in the context of new capabilities that implement business functionality in a timely fashion even as the target infrastructure itself is evolving' [12]. **Model-driven architecture** provides the capabilities of abstraction and encapsulation such that developers can focus on the model itself and ignore all other details, and still reason about it.

From **Observation T.2**, we know that defining fairness requirements needs to be done per model, hence aligning with the idea above. However, from the interviews, we see that certain parts of models might be used to in other models. For example, a model might supply data to multiple different models, making this model a part of the other models' data pipeline. Abstractions should consider this as well.

**M.13** Stakeholders should be able to reason for the model, without having to consider other models in play simultaneously as much as possible

**M.14** Stakeholders should be able to reason for the fairness steps within the parts of the model

### 2. Behavioral Models

[72] describes behaviour models as a precise, abstract descriptions of the intended behaviour of a system. It mentions that due to the nature of behaviour models having solid mathematical foundations, these models can be used to support rigorous analysis and mechanical verification of properties.

Behaviour models combat a limitation of scenario-based requirement generation (e.g. Section 3, in that the former is an exhaustive way of modeling system behavior. While the original aim is to allow for recognition of subtle errors in complex systems ([18] [19]), gaps within behavioral models can promote

exploration and more comprehensive descriptions of the system's behaviors [72].

From this, we can understand that having some form of a concrete record of the system behaviors and actions taken, can allow for deeper thinking about the systems being developed and provide added structure. This is also in line with AR2, wherein showing technical mitigation methods can be considered to support rigorous analysis and mechanical verification. This idea can also facilitate **Observation T.2**, by providing a in-depth record of what actions were taken regarding the system (i.e. the model).

**M.15** Stakeholders should be able to infer which parts of the model fairness has been considered for.

**M.16** Stakeholders should be able to infer which parts of the model fairness has been implemented for.

### 3. Domain descriptors

[71] describes the **Triptych Paradigm within Software Engineering** wherein it describes that "requirements can be expressed properly, the domain of the application must first be reasonably well understood". [5] uses this statement to design requirements engineering system for hospitals. The gist being that, functions within domains can be defined once the specifications of entities is known.

For this project, we can understand entities as the models being developed, and functions as functions that directly enforce fairness, i.e. technical mitigation methods. From this, we understand that the connecting the domains of the models with the the technical mitigation methods **Observation A.4**, might help with expressing the requirements properly.

**M.17** Stakeholders should be able to distinguish between domains to reach to technical mitigation methods

### 4. Property Language

Property languages refers to notations that are derived using automation. This is deemed currently not relevant and out of scope for our project. It can be interesting to explore if the framework should facilitate some form of automatic requirements engineering from the states and behaviors described about a system.

### 5. Notation Semantics

Notation semantics deals with the encoding of various notations as described above. The idea is to keep a standardized format of notations to communicate regarding the model and requirements. We can consider this while developing the framework as well. We should consider **Observation T.1**, and ensure that the semantics are understandable to stakeholders' with different backgrounds.

## Results

### List of Requirements for the framework

Below, we summarize a list of requirements for the framework, in terms of supporting F.R.E.M. :

Requirements ID	Description	Stage in F.R.E.M.
Requirement M.1	Each stakeholder should be able to define their fairness goals	Elicitation
Requirement M.2	Each stakeholder should be able to get an understanding of how realizable the fairness goals are	Modeling
Requirement M.3	Each stakeholder should be able to re-define their goals	Elicitation
Requirement M.4	Each stakeholder can view fairness policies on an institutional level	Elicitation
Requirement M.5	Each stakeholder can view fairness policies on an model level	Elicitation
Requirement M.6	Each stakeholder can reflect on the model level fairness policies	Elicitation
Requirement M.7	Each stakeholder can discuss the model level fairness policies	Elicitation
Requirement M.8	Stakeholders with different backgrounds need to be able to use the framework	Elicitation
Requirement M.9	Stakeholders with different backgrounds need to be able to view each other's viewpoints	Elicitation
Requirement M.10	Stakeholders need to be able to review fairness related scenarios when certain model behaviors occur	Elicitation
Requirement M.11	Stakeholders need to be able to add fairness related scenarios when certain model behaviors occur	Elicitation
Requirement M.12	Stakeholders need to be able to infer non-functional fairness requirements from the framework	Elicitation
Requirement M.13	Stakeholders should be able to reason for the fairness steps of a model	Modeling
Requirement M.14	Stakeholders should be able to reason for the fairness steps within the parts of the model	Modeling
Requirement M.15	Stakeholders should be able to infer which parts of the model fairness has been considered for.	Modeling
Requirement M.16	Stakeholders should be able to infer which parts of the model fairness has been implemented for.	Modeling
Requirement M.17	Stakeholders should be able to distinguish between different ML stages to reach to technical mitigation methods	Modeling

TABLE 3.4: Requirements for designing the framework

### 3.3 Insights for the framework

#### Methodologies, strategies and advice

Methodologies, strategies and advice provide insights on how the requirements derived in Section 3.2 can be realized. We traverse through some methodologies, strategies and advice listed in Table 3.5 for each F.R.E.M stage, to gain interesting insights that can help realize the framework design.

#### F.R.E.M Elicitation

##### 1. Metaphors

[54] studies analogical reasoning for requirements elicitation. Analogical reasoning is a mapping between a base and target description. For example, in requirements engineering, the target description can be heading towards the concrete specification of requirements.

To achieve this analogical reasoning, [54] highlights concepts of the structural mapping theory. These include **Structural Consistency**, **Tiered Identity**, **Systematicity**. The steps involved look as follows :

- (a) Create local matches
- (b) Filter match hypothesis
- (c) Create mappings
- (d) Create candidate inferences
- (e) Evaluate mappings

F.R.E.M Stages	Methodologies, Strategy and Advice
Requirements Elicitation	Nonfunctional requirements Identifying stakeholders, Metaphors, Persona , Inventing requirements, Contextual requirements
Requirements Modeling	RE reference model, Model elaboration, Viewpoints, Patterns, Modeling facilitators, Formalization heuristics, Methodologies

TABLE 3.5: F.R.E.M Elicitation, and Modeling Methodologies, Strategy and Advise noted from [17]

*Remark.* This steps and constraints provide a concrete approach to analogical reasoning that can be used for requirements engineering. We take this into consideration and try to incorporate concepts such as **filtering, mappings, inferences and tiered identity** when designing the framework.

## 2. Stakeholder Identification

[64] promotes stakeholder identification as a part of requirements elicitation. The paper outlines this identification into three main steps, namely :

- (a) Identify all specific roles within the baseline stakeholder group;
- (b) Identify ‘supplier’ stakeholders for each baseline role;
- (c) Identify ‘client’ stakeholders for each baseline role;

*Remark.* From this, we note that designing the framework around baseline stakeholders (i.e. requiring baseline stakeholders to engage with the framework) can be beneficial.

## 3. Contextual Recommendation

[68] proposes a three-layer framework for requirements analysis, wherein one layer is dedicated to analyzing context. They divide this layer in order to explore various aspects where contextual information can be gained. They apply this framework to the assistive technology domain. They find that this domain requires individual and personal requirements to be generated for the success of the product, which can be done via the contextual layer.

*Remark.* We find this approach of a contextual layer, and inferring context from multiple places may be relevant for our project.

From the stakeholder studies performed in Section 3.1.1, we discover that finding model-specific solutions can be vital. This can be seen as heading towards creating *individual* requirements, which [68] shows can be aided using a contextual layer.

## F.R.E.M Modeling

### 1. Viewpoints

[51] introduces ViewPoints, a framework for multi-perspective software development structuring. They state "A ViewPoint, expresses the concerns of a particular stakeholder, such as a development participant or a representative of an area of concern captured by that ViewPoint.". [65] provides a classification scheme for these viewpoints, i.e. what makes a viewpoint unique. The scheme also touches upon the idea of identifying discrepancies and overlaps between viewpoints, aiding in seeing the connection between the viewpoints, and classifying them.

The example applications given in [51] and [65] are of a more *concrete* nature, i.e. the requirements have pre-defined outcomes (e.g. a boolean), making it easier to classify.

**Insight 1** From this, We can aim towards finding some form of overlap and discrepancy between the stakeholder's viewpoints to further connect them.

## 2. Formalization Heuristics

[32] introduce a framework designed to add temporal information to the early requirements engineering phase. To accomplish this, they use dynamic modeling methods and propose *Tropos Grammar*, heuristics to formalize the communication within the framework. This gives consistency to the dynamic modeling, they try to accomplish.

### Insight 2

From this, we understand that the building blocks of the framework, and their interaction can benefit when supplemented with a concrete definition.

## 3. Patterns

[25] states that most industrial requirements engineering is done in natural language. It aims to remove the ambiguity and imprecision that may arise from this practice, by proposing *Natural Language Patterns*. These *Natural Language Patterns* are quite interesting, as they map various common terms used such as *then, until* or *as long as* and set a pre-defined meaning to them.

### Insight 3

From this, we understand that we can look towards minimizing ambiguity of natural language requirements. We can think about prompting the stakeholders to automatically categorize some of their requirements. This can be done by issuing constraints, or a pre-defined matrix to be filled.

## 4. Model Elaboration

[74] meticulously motivates the usage of implied scenarios. We came across the concept of *implied scenarios* in Section 1.

To recall, [74] defines implied scenarios as "Implied scenarios identify gaps in scenario-based specifications that arise from specifying the global behavior of a system that can be implemented component-wise. They are the result of a mismatch between the behavioral and architectural aspects of scenario-based specifications."

The study goes on to show that the addition of implied scenarios can prompt stakeholders' to think deeper into the requirements coverage.

### Insight 4

Insight ID	Description
Insight 1	From this, We can aim towards finding some form of overlap and discrepancy between the stakeholder's viewpoints to further connect them.
Insight 2	From this, we understand that the building blocks of the framework, and their interaction can benefit when supplemented with a concrete definition.
Insight 3	From this, we understand that we can look towards minimizing ambiguity of natural language requirements. We can think about prompting the stakeholders to automatically categorize some of their requirements. This can be done by issuing constraints, or a pre-defined matrix to be filled.
Insight 4	From this, we see that allowance some form of communication of scenarios through out the development and production phase of a model, can help identify the gaps in contextually derived requirements.
Insight 5	From this, we understand that capturing behaviors through time can be relevant to ensuring a sustainable way of requirements engineering. We also note that the identification of various behaviors at the start, can already aid this.

TABLE 3.6: Insights for designing the framework

From this, we see that allowance some form of communication of scenarios through out the development and production phase of a model, can help identify the gaps in contextually derived requirements.

#### 5. RE reference model

[76] argues that the goal-oriented requirements specification can lead to idealistic and unrealistic requirements being generated, that to not consider exception behaviours of the system. Thereby, resulting in the system lacking robustness, poor performance or failures. To address this, they propose a two layer framework.

The first layer's objective is to specify goals and objectives, while the second layer can be used to re-specify these goals. The key here, is to allow for handling of exceptions at requirements engineering time so that the re-specification becomes easier.

[37] (re-)formulating the requirements engineering model introduced in [36], to include the consideration of behavioral changes over time, and how that can affect the requirements. This leads to more relevant requirements being specified, and new behavioral requirements begin identified for different times.

**Insight 5** From this, we understand that capturing behaviors through time can be relevant to ensuring a sustainable way of requirements engineering. We also note that the identification of various behaviors at the start, can already aid this.

## Results

### List of Insights for the framework

Below, we summarize a list of insights for the framework, in terms of supporting F.R.E.M. :

## Chapter 4

# Multi-Level Fairness Framework

In this chapter, we tackle Research Question 3b. We introduce the **Multi-Level Fairness Framework (M.L.F.F.)**. The completed framework is displayed in Figure 4.2. We wish to tackle :

1. **Objectives : Designing the M.L.F.F.** : We design the framework by determining its components, mechanisms and actions based on the overview in Chapter 2, and reflection on Chapter 3's findings.  
**Usage of the M.L.F.F.** : We elaborate on the usage of the framework, to envision how the components, mechanisms and actions work together, and a scenario where the stakeholders interact with the framework.
2. **Methodology** : In chapter 2, we gained an overview of state-of-the-art practices to perform F.R.E.M. In Chapter 3, we discovered **Institutional Challenges to address, Requirements to support F.R.E.M. and Insights to support F.R.E.M.**, we derive the M.L.F.F.
3. **Results** : We present the framework for fairness requirements elicitation and modeling (F.R.E.M), coined the Multi-Level Fairness Framework, by walking through its specific characteristics and adding context with a use case scenario.

### 4.1 Objectives

In this section, we will design the framework. The framework will consist of components, mechanisms and actions.

We define these terms as follows :

1. A component is an entity within the framework. The framework should consist of at least one component.
2. A mechanism has a particular function it accomplishes, to improve the interaction or usage of the components themselves. Mechanisms between components are not mandatory.
3. An actions are basic tasks that can be performed by the stakeholders, that involve interacting with the components or mechanisms. For each component, at least one action should be defined.

For example, a component can be *Fairness Definitions* and *Fairness Metrics*. A mechanism between these two components can be *Mappings*, which maps *Fairness Definitions* to *Fairness Metrics*, improving the interaction between these two components. The relevant actions can consist of *Reviewing*, where the stakeholder can review the *Fairness Metric* based on the *Fairness Definition* it is mapped from.

## 4.2 Components of the framework

We map the requirements in Table 3.4, to the existing entities discovered in Figure 2.2, refining them and adding additional entities if required. We convert these entities into the components of the framework, by formally defining them. We divide the type into Fairness Requirements Elicitation Components, in light of Requirements M.1 to M.12 in Table 3.4, and Fairness Requirements Modeling Components, in light of Requirements M.13 to M.17.

### Fairness Requirements Elicitation Components

For Model stakeholders to be exposed to multiple fairness notions and definitions, we use the three fairness entities discovered in Chapter 2, namely Social Fairness Notions, Fairness Definitions and Fairness Metrics. These three entities encapsulate fairness in different terms, allowing for comparisons between notions to be made, and allowing stakeholders to understand fairness on a philosophical level to a statistical level. Below, the formal components (accompanied by their definition, an example and requirements mapping to Table 3.4) are stated :

#### Component : Social Fairness Notions

**Definition 4.2.1** (Social Fairness Notions). A social fairness notion describes a concept (abstract or concrete) that eludes to what is considered fair. A notion or definition is categorized as *social*, if it does not overlap with the component of *Fairness Definitions* or *Fairness Metrics* in the M.L.F.F.

In general, social refers to the notion not being popular, or having been considered **directly** in the community of *Machine Learning* (or related) fields. For example, these fields may belong to *Liberal Arts*, or *Law*.

#### Component : Fairness Definitions

**Definition 4.2.2** (Fairness Definitions). A fairness definition is a criteria, measure or concept that represents the meaning of fairness. A fairness definition must be directly mapped to a particular technical mitigation method, or fairness metric.

In general, fairness definitions are definitions used within the *Fairness in Machine Learning* field. Thus, we expect there is some form of algorithmic achieve ability in these definitions.

#### Component : Fairness Metrics

**Definition 4.2.3** (Fairness Metrics). A fairness metric states the mathematical execution of a fairness definition.

For this project, we will use the fairness metrics defined in Section 2.3, focusing on binary classification tasks.

For Model stakeholders to review and add contextual information, all components that provide social context satisfy this requirement. This includes the components of social fairness notions, institutional principles (introduced further ahead) and trade-offs (introduced further ahead).

We still want to facilitate specific contextual information to the model, hence we introduce the component of a case study. Case Studies allow for an insight on past



fairness cases that have occurred in regards to the model, or similar model type. Case Studies is a set of information inputted by the stakeholders, to be able to consider and refer to.

### Component : Case Studies

**Definition 4.2.4** (Case Studies). A case study is a review-able case that stakeholder deems relevant to be communicated in for the current model in development. A case study can be comprised of :

1. Name - Name of case study
2. Date - Date occurred (time references can elude to social context)
3. Domain of Application - Domain of application (e.g. Credit Risk, HR)
4. Domain of Model - Type of Model (e.g. Entity-Recognition Model, Binary Classification)
5. Description - A description of what occurred
6. Mitigation Methods Applied - Inclusion of mitigation methods that were used
7. Notable Actions - Actions, related to fairness, that led to impact-ful consequences, be it negative or positive

The relevance of a case study is up to the stakeholder. Stakeholders can add cases related to the particular domain (application or model). If sufficient information is not available, stakeholders can flock to including cases from related domains. Alternatively, stakeholders can decide to deliberately include cases from varying domains, or models.

To allow the Model stakeholders to add information relevant to the institutional policies regarding fairness, we introduce the component of *Institution Wide Principles*, which can contain all institution relevant principles, approaches and guidelines regarding fairness.

### Component : Institution Wide Principles

**Definition 4.2.5** (Institution Wide Principles). Encodes all over-arching principles, or complication and legal notions that the institution wishes to be considered when designing the fairness requirements.

We argue that any encoding that needs to be done on a model level, can already manifest itself into the fairness notions being selected of that model. Hence, Institution Wide Principles remain the same for all models.

## Fairness Requirements Modeling Components

### Component : Trade-offs

**Definition 4.2.6** (Trade-offs). A trade-off consists of strictly two concepts associated with algorithmic and fairness notion based. This two concepts must obverse a contradictory nature in achieve-ability, wherein a compromise may arise. A trade off consists of :

1. Concept 1 - The name of the first concept
2. Concept 2 - The name of the second concept
3. Discussion - Discussion on the trade-off between Concept 1 and Concept 2

Here, the discussion is textual. Intuitively, having a quantitative measure (e.g. likert scale) for trade-offs can be helpful. We find that currently there is a gap in literature wherein quantification of trade-offs in fairness does not exist. Hence, we refrain from adding it to our definition.

To prompt Model stakeholders to research in-depth regarding mitigation method, we wish to expose them to existing and new literature regarding technical (un)fairness mitigation methods. We discovered the various mitigation methods in Chapter 2, and how they differed in terms of application, and the fairness definitions/metrics that were targeted. Hence, we introduce the component of *Technical Mitigation Methods*, that encapsulates a method, with the conditions of its application and original literature paper (if present). A mitigation method also consists of the conditional statement in Fairness Metrics (or definitions) associated, so that mappings to fairness metrics and definitions can be made more structurally. This can be eliminated, should the stakeholder desire to go with a more flexible approach.

Furthermore, to prompt in depth research, we can utilize the specification of the Machine Learning pipeline (stages and activities), that we discovered in Chapter 2. This can prompt the stakeholders to look for mitigation methods relevant to their activities, allowing the focus of fairness to be distributed throughout the pipeline.

### **Component : Technical Mitigation - Mitigation Methods**

**Definition 4.2.7** (Mitigation Methods). A mitigation method is a technical activity (or a series of activities) that try to accommodate a particular fairness metric or fairness definitions. A mitigation method consists of :

1. Domain of model - Which domain (of model) is it relevant for (e.g. NLP, Computer Vision, Classification)
2. Method Description - Description of the method, can be a reference to the literature paper, code or toolbox.
3. Method Evaluation - Description of the criteria to be evaluated in order to measure the extent of *fairness* achievement
4. Conditions of application - Conditions (algorithmic, data-wise, problem-type) that may restrict the use of this method
5. Fairness Metrics (or definitions) associated - Which fairness metric does it try to accommodate. If no fairness metric is specified, then one may indicate the fairness definition.

### **Component : ML Pipeline Stages**

**Definition 4.2.8** (Stages). A stage belongs to the set of stages that are considered as common steps when designing a ML Pipeline. A stage consists of :

1. Name - The name of the stage

2. Description - Brief description of the goals of this stage.
3. Activity(s) - List of all activities that help accomplish the goal(s) for this stage. List is only comprised of activities in component

For this project, we will use the stages defined in Table 2.1.

### Component : ML Pipeline Activities

**Definition 4.2.9** (Activities). An activity is a technical task that is conducted, within a particular stage from the component *Stage* of the M.L.F.F. An activity consists of :

1. Name - The name of the activity
2. Description - Brief description of the goals of the task that need to be done.
3. Stage(s) - The list of stages from M.L.F.F., whose goal(s) are aligned with the task goal(s).

The stakeholders can add in the stages of the Model.

To provide the Model stakeholders with some form of systematic way of arranging mitigation methods and referring to them, we can utilize the Machine Learning Pipeline entities we discovered in Chapter 2, namely stages and activities to organize the mitigation methods. The relevant components for this, are ML Pipeline Stages, and ML Pipeline Activities.

To allow Model stakeholders to access information relevant to a particular model, we introduce the component of Models, which encapsulates particular information regarding the model. This component can then be connected to instances of the other components, hence allowing stakeholder to extract fairness information relevant to the model as well.

Below, we define the component of Models :

### Component : Models

**Definition 4.2.10** (Models). A model is the machine learning part of a *technical project* being done within the institution. A model should include (only) one primary *training* task. A model comprises of :

1. Name - Name of model
2. Description - Brief description of the goals of the model
3. Domain of model - The technical domain the model, or the goal falls into. (e.g. NLP, CV, Classification)
4. Domain of application - The domain in which the model be used in (e.g. Customer Dialog, HR, Workforce Prediction).
5. Pipeline - A list of stages from Component *Stages* in M.L.F.F., that make up the ML pipeline of the model.

We constraint the model definition one primary *training* task, as this provides some way of segmenting between sequentially used models. This constraint can be removed or replaced by an institutional convention as well, should the stakeholders wish this.

### 4.3 Actions for the framework

Till now, we have outlined the components of the M.L.F.F. and what they entail. We have yet to define the protocol required for interacting with these M.L.F.F. components. The study of the notations provided us with requirements for facilitating F.R.E.M Elicitation and F.R.E.M Modeling for the stakeholders.

In this section, we present the actions inferred and motivate them using the requirements (Table 3.4).

**Definition 4.3.1 (Action : Review).** Only read the information present, indicated or added

**Definition 4.3.2 (Action : Indicate).** Indicate opinions on pre-defined information in the framework. Pre-defined information can only be added or changed by the moderator of the framework, (not the stakeholders).

**Definition 4.3.3 (Action : Re-adjust).** Change indicated information

**Definition 4.3.4 (Action : Add).** Can add information to authorized components. Unauthorized components can only be added or changed by the moderator of the framework, (not the stakeholders).

We motivate that these actions and their combinations (set of actions) can facilitate all requirements mentioned in Table 3.4. This property of *set of actions* is crucial to the M.L.F.F. If we consider the nature of the observations identified in Chapter 3 stating that the Level of Awareness and Level of Transparency should be improved to allow for fairness requirements engineering, we understand that different use-cases of the M.L.F.F. can call for different set of actions to be used.

For example, Requirement M.1. can be realized by multiple set of actions. If the stakeholder does not have sufficient awareness regarding the fairness goals, then their course of action can be *Review, Indicate*, whilst a stakeholder that has sufficient awareness may only require the action of *Indicate* to fulfill Requirement M.1. Let us take another requirement, Requirement M.6., where stakeholders can *reflect* the model level policies. Here, for example, if the stakeholder is a Model Developer, than this reflection may only consist of the action of *Review*. If the stakeholder is a Model Owner, *reflection* can also consist of *Review, Indicate*.

Similarly, for example, for Requirement M.14 and M.15, the stakeholders can consist of two model developers that are working on different stages of the model. Then, the model developer can have the action *Review, Add, Indicate* for its responsible model stage, whilst only require the action of *Review* for the other model stages.

With this, we show that the protocol for actions is purposely kept flexible as in Chapter 3, we noted that consideration for different use-cases for the M.L.F.F. needs to be present.

### 4.4 Mechanisms for the M.L.F.F.

We look towards the insights, to see any mechanisms that need to be introduced to facilitate F.R.E.M Elicitation and F.R.E.M Modeling. To infer these mechanisms, we look towards the insights outlined in Table 3.6, that encapsulate the possible Methodologies, Strategy and Advise given for F.R.E.M Elicitation and F.R.E.M Modeling. These mechanisms will allow us to further strengthen the components and actions of the M.L.F.F., as it will allow for research findings to be considered alongside findings from Chapter 3.

To clarify, the difference between a mechanism and a component is that a mechanism has a particular function it accomplishes, to improve the interaction or usage of the components themselves. The difference between an action and a mechanism is that, actions are basic tasks that can be commenced by the stakeholders, whereas a mechanism can consist of actions that then help accomplish a particular motive.

### Mechanism : Perspectives

**Definition 4.4.1** (Perspectives). A perspective captures the viewpoint of a particular stakeholder, which can then be referred to by other stakeholders.

A referable perspective is captured against a pre-defined matrix, where the pre-defined matrix should be comprised of capturing viewpoints on either fairness notions or trade-offs.

**Function:** Provide a way to further connect the stakeholders by outlining their overlapping and diverging perspectives (**Insight 1**). By allowing stakeholders to view each other's viewpoints on a pre-defined matrix (in our case, it is the matrices involved with the component of **Social fairness notions** and **Trade-offs**), we can add concrete-ness in this communication (**Insight 2**).

### Mechanism : Mappings

**Definition 4.4.2** (Mappings). A mapping can occur between any two (or more) objects within a component.

**Function:** We want to introduce pathways connecting different components to capture constraints and possibilities (**Insight 2 and 3**) of traversing through different components. We can do this, in the form of mappings between components.

From our background study, we already come across various mappings that occur within literature, for example Aequitas [1], which provides a flow diagram to decide which fairness definition may be suited, or consider the mapping between inequalities and fairness philosophies given by [46]. Even, looking towards technical mitigation methods, [29] maps basic stages of the Machine Learning model to technical (un-)fairness mitigation methods for binary classification. These mappings, when populated over time (as research is conducted) can provide pathways from components.

### Mechanism : Dilemmas

**Definition 4.4.3.** Hence, we can introduce the mechanism of *Dilemmas* which provide an abstract situation that deliberately puts stakeholders into *unfair* context, with the intention to prompt deeper thinking regarding unfair situations.

**Function:**

Dilemmas can be added to prompt deeper thinking for many components. Dilemmas can be seen similar to case studies, however case studies are recorded and occurred cases, whilst dilemmas can take up an abstract form. One finding from our stakeholder study conducted within 3, is that the study itself prompted the stakeholders to think deeper about the fairness scenarios (**Insight 4**). As these scenarios are hypothetical in nature, one can create dilemmas to fit the societal expectations as time progresses (Dilemmas can be added to prompt deeper thinking for many components.).

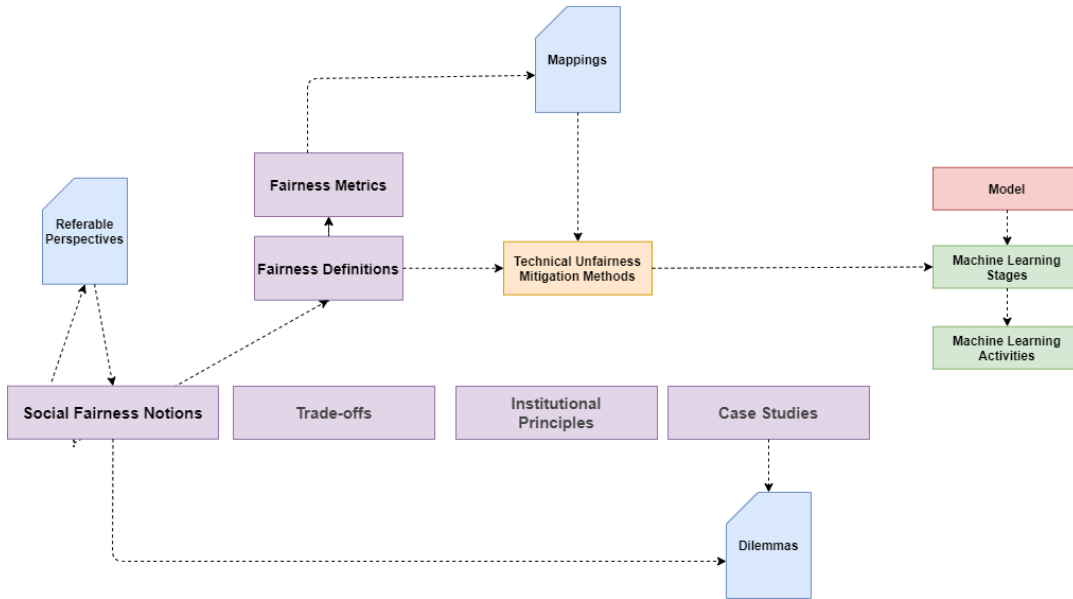


FIGURE 4.1: Framework with Sub-Mechanisms

For example, one dilemma we discover during the stakeholder studies, was in Observation A.1, where the translation of a confusion matrix metric to the context (so creating a dilemma by having the stakeholders select preferences between false negatives or false positives within a particular context) inferred different answers from the stakeholders, and revealed underlying perspectives.

However, one glaring limitation of this mechanism is that it could prompt bias within stakeholders if the dilemmas are themselves biased. Hence, we propose that dilemmas are only introduced by stakeholders within the institution after deliberation and thought.

## 4.5 Characteristics of the framework

Above, we derive and formalize the blocks that build up the Multi-Level Fairness Framework. In this section, we want to allow the reader to build an understanding of how the components, mechanisms and actions (derived above) work together. Therefore, we touch upon three main characteristics of the M.L.F.F. (which we refer to as a nature of the framework) and walk through them, to build an intuition on how the entirety of the framework works. We tackle the following *natures* :

1. Multi-Level Nature : Here, we explain and motivate the *multi-level* nature of the framework and visualize how it fits in with the components, mechanisms and actions.
2. Information Nature : Here, we explain an example of the information that can be contained within the framework with the help of a dummy use-case of Credit Risk Modeling. This allows the reader to build an intuition on the type of information each component may contain, for a particular model.
3. Role-Based Nature : Here, we walk through an example of the different stakeholders interacting with the framework with the help of a dummy use-case of Credit Risk Modeling. This allows the reader to build an intuition on how the framework is designed for different stakeholders.

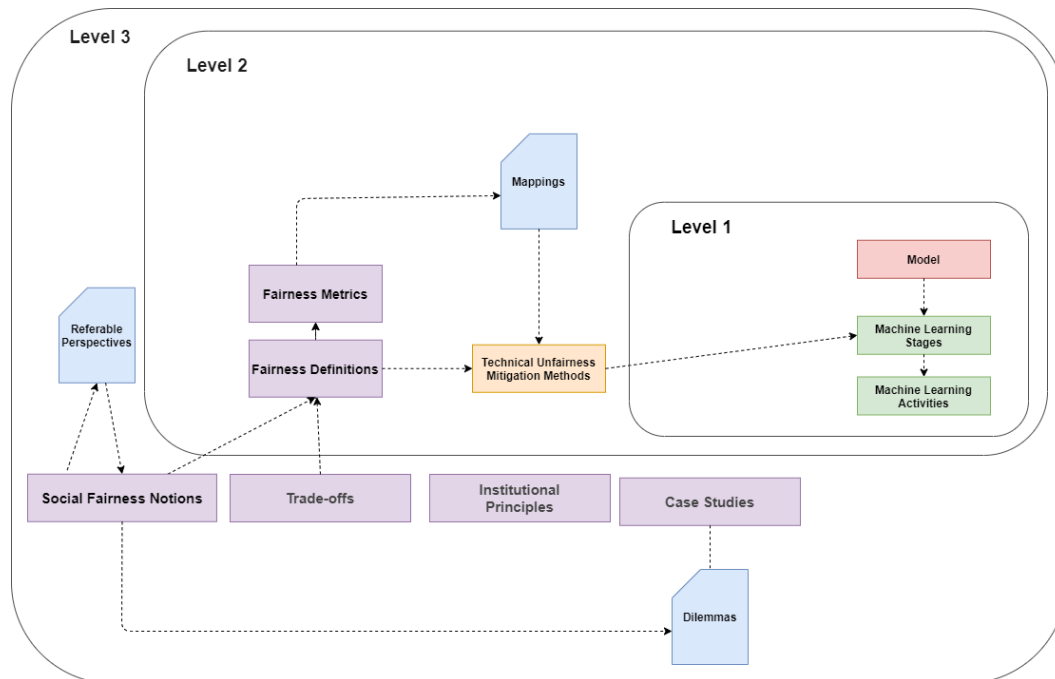


FIGURE 4.2: Multi-Level Fairness Framework

4. Challenge-Nature : Here, we highlight which challenges, in terms of Consideration, Specification and Communication we expect components and mechanisms to address.

#### 4.5.1 Multi-Level Nature

In Figure 4.2, Level 1 is the base level where the stakeholders can define the models being developed and ML Pipeline associated with this model. We expect this information to be available for immediate use.

Level 2 maps the ML Pipeline stages to the technical mitigation. This always the stakeholders to become aware of mitigation methods and metrics that can be applied to each stage. This can be done, without necessarily deciding on the fairness definitions that are applicable for the model.

Level 2 maps fairness definitions to technical mitigation. If the stakeholders have an idea of which technical mitigation method they would like to use, then they can directly infer the relevant mitigation methods. If can be done without necessarily having the social context, or social fairness notions decided upon.

Level 3 consists of social context and social fairness notions that can then be added through time. Relevant fairness definitions can be deliberated in-depth by referring to social context or social fairness notions.

#### 4.5.2 Information Nature

We provide a high-level overview of what each of these components entail, with an example of a credit-risk model.

Let us assume the binary classification credit-risk model built in 2019 which being updated in 2021 within the bank and there is a high-level task of ensuring fairness within the model. Then, for the component :

1. **Model** will be Credit Risk Model v2021

2. *ML Pipeline Stages* will be the main stages of the ML pipeline, for example *Data Collection*, *Feature Engineering*, *Model Training* and *Model Evaluation*
3. **Case Studies** can be as follows : "In Credit-Risk Model v2019, there was a detection of proxy features within the model that eluded to the gender of the subject"
4. **Fairness Definition** can be set as *Equal Odds* where the intuition is as follows "Suppose Group A and Group B contain two features, where one feature is sensitive (i.e. you want to make sure you are being fair towards this feature, e.g. Group A can be Male, and Group B can be female). Among applicants who are creditworthy and would have repaid their loans, both Group A and Group B applicants should have similar rate of their loans being approved"
5. **Fairness Metric** will be set as *Equal Odds* as well, "An algorithm is considered to be fair under equal odds if TPR and FPR are considered simultaneously. Formula is  $Pr(\hat{y} = 1|y = 1g_i) = Pr(\hat{y} = 1|y = 1g_j)Pr(\hat{y} = 1|y = 0g_i) = Pr(\hat{y} = 1|y = 0g_j)$ "
6. **Social Fairness Notions** can consists of : Characteristics pertaining to the *socio-economic* and *talent inequalities* are allowed to be used. Characteristics pertaining to *natural inequalities* are not allowed to use. The inequalities are described in Table A.1. Or for example, the fairness philosophies behind Equal Odds can be indicated so : Fair Equality of Opportunity is a relevant fairness philosophy for this model. Equal Odds is a fairness definition that complies with this fairness philosophy.
7. **ML Pipeline Activities** will entail granular activities present within these stages. One can indicate a general activity or activities specifically related to fairness. So, the activity of *Checking confusion matrix* can be an activity listed under *Model Evaluation*. Within this activity, the fairness steps can be listed as *Checking TPR and FPR for gender* and *Checking proxies for gender*.
8. **Trade-offs** will consist of trade-offs to be recorded on multiple dimensions which are shown in Table A.3. For example, the dimension of *internal data quality* can have the following trade-off recorded : implementing Equal Odds for feature *gender* requires a certain distribution of data with the feature of *gender*. Getting access to this data needs to be approved by the Data Protection Officer, to ensure Equal Odds fairness for two groups within the feature of *gender*.

### 4.5.3 Role-Based Nature

We use an **activity diagram** in combination with the example, to show how the framework functions, and what each level provides to show the role-based nature of the framework (Figure 4.3, 4.4 and 4.5). An activity diagram shows how activities are coordinated for the use of the Multi-Level Fairness Framework..

To recall briefly, the stakeholders using the M.L.F.F. are identified as the model owner, domain expert and model developer. Model owners hold responsibility for formulating and communicating fairness expectations to model developers. This can require them to consider various aspects such as contextual information and societal expectations, and institution expectations. Both model owners and domain experts are in charge of providing information on the model application to the model developer. Model developers are in charge of technical responsibility of executing



the requirements set by the model owners. Technical responsibility can entail communicating technical feasibility and technical executions.

### Use-Case Description

There is a credit-loan risk model being built, with the following stakeholders involved :

1. Two Model Developers :  $Developer_{C1}$ ,  $Developer_{C2}$  wherein  $Developer_{C1}$  is in charge of the building the algorithmic model and  $Developer_{C2}$  is in charge of data collection for the model.
2. Model Owner
3. Two Domain Experts :  $DomainExpert_1$  and  $DomainExpert_2$  wherein  $DomainExpert_1$  has expertise on customer relationships for loan applications and  $DomainExpert_2$  has expertise on the financial thresholds of providing a loan.

*Remark.* This section focuses on the interaction of components, actions and mechanisms. If the reader wishes to recall the exact details of the components, actions or mechanisms, they can refer back to Section 4.2, 4.3 and 4.4.

### Level 1

Level 1 of the usage is shown in Figure 4.3. The  $ModelDeveloper_{C1}$  adds the model  $CreditRiskModel$ . This model instance contains the name, description, domain of model and domain of application of the model.  $ModelDeveloper_{C1}$ , then goes ahead and indicates the relevant stages for this model. Since  $ModelDeveloper_{C1}$  is in charge of the algorithmic model training and evaluations, the stages indicated are **Model Training**, and **Model Evaluation**, with further addition of activities. In Figure 4.3, we show example activities of **Optimization** and **Cross-Validation**.

Since  $ModelDeveloper_{C1}$  has already created an instance of  $CreditRiskModel$ ,  $ModelDeveloper_{C2}$  can indicate this model, and proceed to indicate their relevant stage and add activities to this. In Figure 4.3,  $ModelDeveloper_{C2}$  indicates the stage of **Data Collection** and adds the activity of **Labeling**.

This concludes the interactions occurring at Level 1.

### Level 2

Moving on to Level 2, the model owner can indicate the  $CreditRiskModel$ . The model owner is prompted towards indicating relevant fairness definitions for the model. Here, the model owner can use existing mappings of Aequitos [2] (for example), alongside the descriptions of the fairness definitions to narrow the relevant fairness definitions. In Figure 4.4, the model owner indicates two interesting fairness definitions for the model, namely **Demographic Parity** and **Equality of Opportunity**. The model owner can then review the required factors for these two fairness definitions. The model owner adds the relevant **protected attributes** for Demographic Parity and **protected attributes, target variable, utility and benefit** for Equality of Opportunity. In Figure 4.4, the  $DomainExpert_2$  indicates the model, and can review the relevant fairness definitions and add in other factors that they may possess knowledge over (for example the **prediction probability score**, and **prediction threshold values** from the business perspective). This communication of the  $DomainExpert_2$

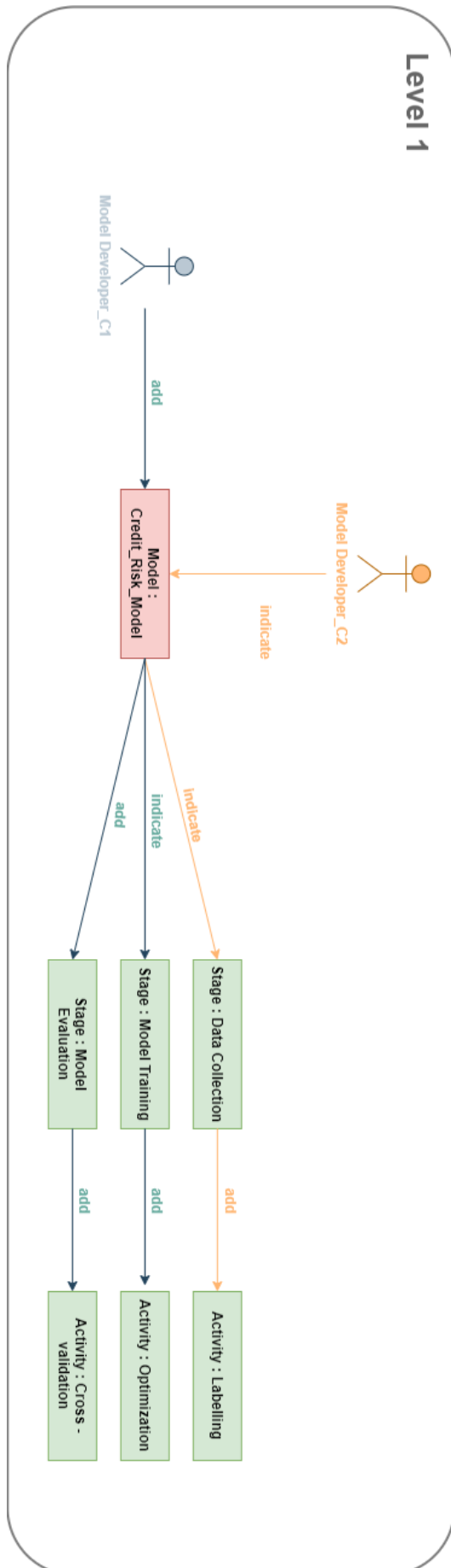


FIGURE 4.3: Model Owner Engagement with the Framework

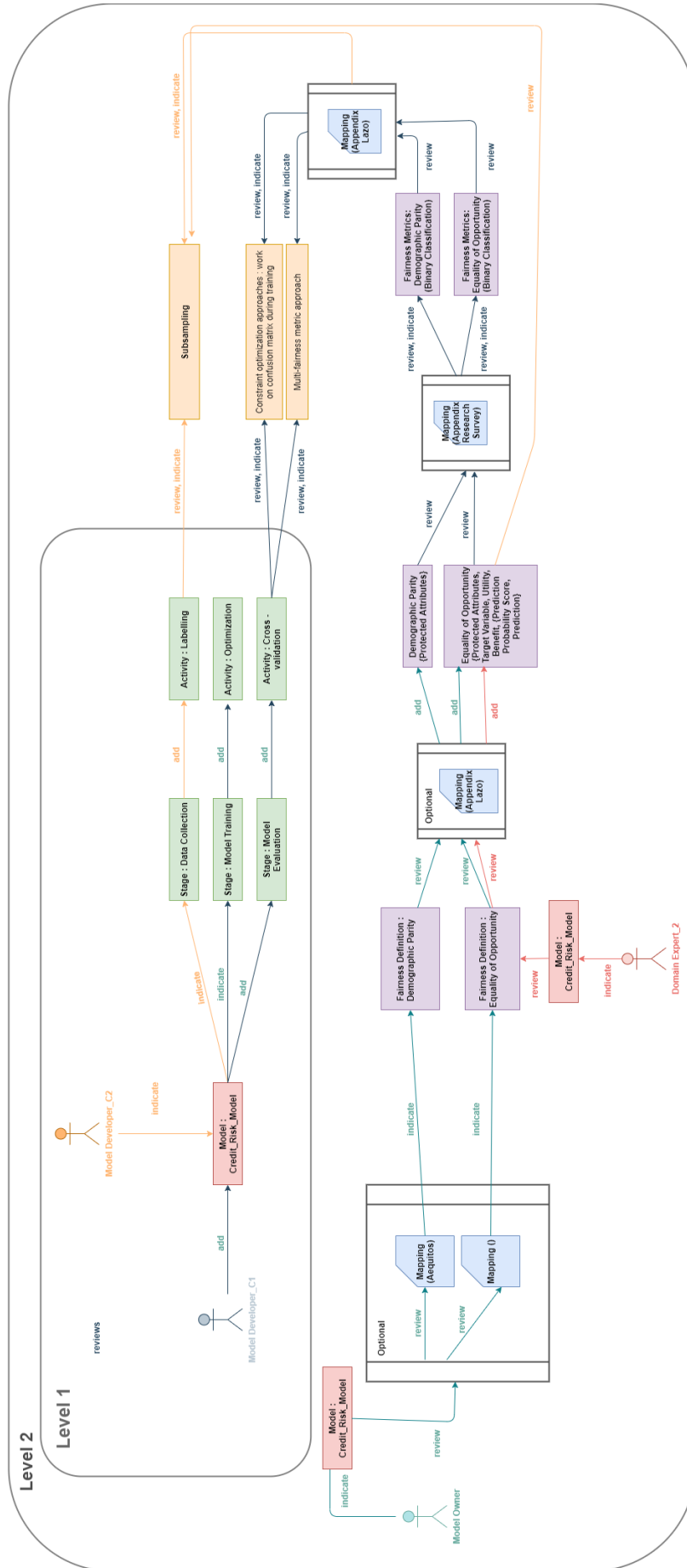


FIGURE 4.4: Model Owner Engagement with the Framework

is assumed to be done out of the framework, when the *model<sub>owner</sub>* realizes that this information is more suited to be filled in by a domain expert.

After the factors and relevant fairness definitions are indicated, the *ModelDeveloper<sub>C1</sub>* and *ModelDeveloper<sub>C2</sub>* can review these definitions, and via a mapping, can review and indicate the relevant fairness metrics that need to be applied.

The model developers can then review whether any mitigation methods exist for these particular metrics/definitions, per model stage. For example, in Figure 4.4, *ModelDeveloper<sub>C1</sub>* considers the technical mitigation methods of **Constraint Optimization Approaches** and **Multi-Fairness Metric Approach** and reviews and indicates it within the stage Model Evaluation, and the activity Cross-Validation. *ModelDeveloper<sub>C2</sub>*, reviews and indicates the technical mitigation method of **Sub-sampling** for the model stage of **Data Collection** and activity **Labeling**. This concludes the main interactions taking place within Level 2.

### Level 3

In Level 3, the model owner indicates the model of *CreditRiskModel*, and can indicate social fairness notions. In Figure 4.5, the model owner indicates the **Fairness Dimensions** (Table A.3), **Inequality Levels** (Table A.1) for the *CreditRiskModel*.

To gain knowledge of the domain expert, *DomainExpert<sub>1</sub>* indicates the Fairness Dimensions, Inequality and Bias Levels as well.

Both the Model Owner, and *DomainExpert<sub>1</sub>* have the option to review each other's perspectives via **Referable Perspectives** and to think deeper via **Dilemmas**.

The model owner can then re-adjust the relevant definitions (specified in Level 2) if necessary.

In Level 3, we also see that after reflecting on the **Institutional Principles** and **Trade-offs**, and having knowledge regarding the relevant fairness metrics and mitigation methods (from Level 2), the *ModelDeveloper<sub>C1</sub>* adds a trade-off between two biases. The *ModelOwner* can review this trade-off and based on the conclusion, interactions within Level 2 can then be re-adjusted. For example, maybe the *ModelDeveloper<sub>C1</sub>* needs to find a new mitigation method, or *ModelOwner* may reconsider fairness definition (and subgroups).

With this, we show an example use-case of the M.L.F.F. to provide the reader with a complete overview. Of course, this use-case is not a complete encapsulation of all possible scenarios, but it presents an understanding on the usage of the Multi-Level Fairness Framework.

#### 4.5.4 Challenge-Nature

In the Table 4.1, we give an overview of the framework components and mechanisms, and the institutional challenges we expect them to address.

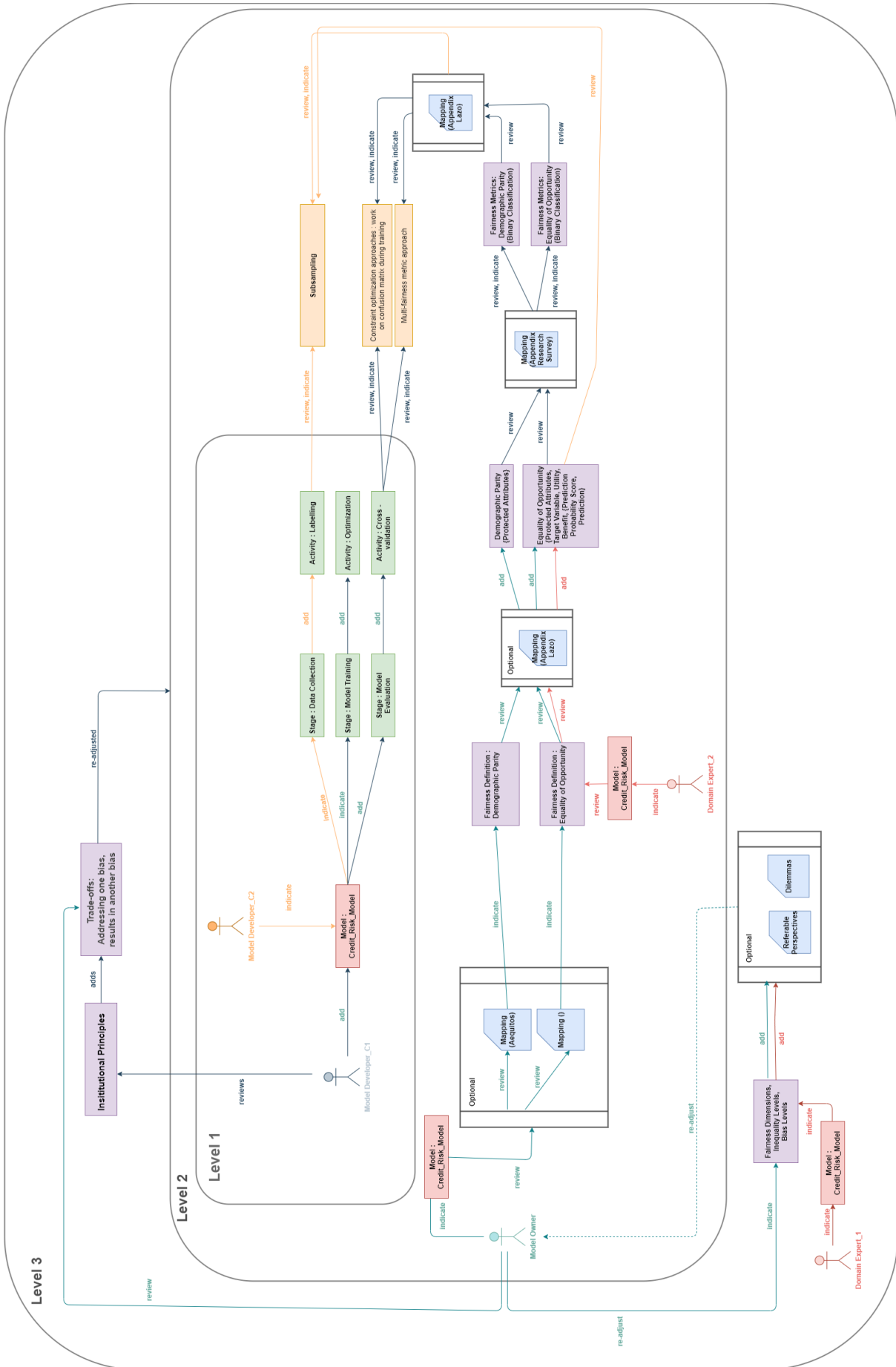


FIGURE 4.5: Model Owner Engagement with the Framework

M.L.F.F. Part	Challenge
Social Fairness Notions	Consideration (of different fairness notions that display variations of what "fairness" means)
Institutional Principles	Consideration (of institutional ideologies, and visions)
Case Studies	Consideration (of fairness within social context on the ML model)
Dilemmas	Consideration (of fairness within social context on the ML model)
Trade-offs	Consideration, Communication, Specification (of decisions being made, or to be made between different dimensions that involve trade-offs between fairness for particular groups, business acumen, model performance, resources and so forth)
Fairness Definitions	Specification (of which fairness definitions are mandatory, to experiment or not applicable)
Fairness Metrics	Specification (of what the empirical translation of a fairness definition means for a particular ML task (e.g. binary classification))
Technical Mitigation Methods	Consideration (of various mitigation methods that can be applied at various stages of the ML model pipeline)
Machine Learning Stages	Consideration, Communication (of what fairness steps are recorded, or need to be recorded within a particular stage. And an overview of fairness steps taken in different stages, which can be reflected upon)
Machine Learning Activities	Consideration, Communication, Specification (of which activities are being performed, and consideration and specification on which fairness steps need to be taken)
Mappings	Communication, Specification (on which fairness inequalities link to which fairness philosophies then fairness definitions. And which fairness definitions link to which fairness metrics. So, only stakeholders with corresponding backgrounds are exposed to their particular entities.)
Perspectives	Consideration (on what other perspectives are regarding fairness for the ML model)

TABLE 4.1: Targeted Challenges of the components

## Chapter 5

# Evaluation

In this chapter, we tackle Research Question 4. We obtain insights to evaluation the framework.

1. **Objectives: Evaluation of the M.L.F.F:** To what extent does M.L.F.F. address consideration, specification and communication to perform F.R.E.M. to be used on further iteration development of the M.L.F.F.
2. **Methodology:** We build a prototype and perform a qualitative evaluation stakeholder studies that consist of case studies and semi-structured interviews with the stakeholders. We design the case study to experiment and gain insights on the M.L.F.F. usage.
3. **Result:**
  - (a) **Findings :** The discussion on the findings that are extracted from the evaluation stakeholder studies, and what they indicate in terms of addressing the challenges in F.R.E.M., and potential improvements suggested or inferred from the stakeholders present in the evaluation stakeholder studies.

### 5.1 Objectives

Evaluation is the process of comparing the effects of the use of the artifact with selected criteria to conclude whether it is satisfactory [verschuren and hartog, 2005]. To evaluate the prototype, we first identify the objectives we wish to serve for the next iteration of improving the framework (and prototype).

In this project, we build on a relatively new field, where we infer potential practices from literature studies, investigate institutional challenges and build a framework. Our goal for this evaluation is primarily to get an idea of how the framework components address F.R.E.M, and whether the components tackle Consideration, Specification and Communication to perform F.R.E.M.

We refrain from directly evaluating whether the framework facilitates F.R.E.M. as this would require us to measure whether the fairness requirements elicited and modeled are actually correct for a particular model. Since our problem statement is to "support F.R.E.M", we are interested in seeing whether the framework can address challenges within F.R.E.M, thereby supporting the stakeholders in performing F.R.E.M.

Evaluation Study	Stakeholder	Simulated Roles	Evaluation Role
Study A (Appendix C)		Model Owner	Two Model Developers
Study B (Appendix C)		Model Developer	Two Model Owners

TABLE 5.1: Stakeholder Study Setup

## 5.2 Methodology

### Prototype

To evaluate M.L.F.F. versus status quo, we build a digital prototype of the framework. We use Streamlit [67], an open-source app framework based in Python, and Cloud Firebase [31], a cloud-hosted, NoSQL database to create the front-end and back-end of the prototype. Below, we show a few visualizations of the prototype. In Appendix ??, the reader can find each component mapped to a visualization within the prototype.

*Remark.* Mappings are including within the back-end of the prototype, where for example, if a stakeholder indicated a fairness definition, then the corresponding fairness metric is automatically shown. These mappings were inferred from Chapter 2, and are implemented with queries.

We exclude the components of *Institutional Principles*, and *Perspectives* as these components are evaluated through the semi-structured interviews, and asking questions on it. The motivation for this, is that a) it is not clear what the institutional principles may entail (the depth of knowledge) and the same applied for "Perspectives" as it is not clear to what depth the discussion can be. We can simulate these aspects, but these aspects may differ for every person using the prototype. For example, one person may go into a lot of detail regarding their perspective, whilst another may not. So, here our objective is to understand whether these components can actually be beneficial (rather than evaluating what these components should be designed like). Interestingly, we will find that the stakeholders themselves notice the absence of these components.

### Design of the evaluation studies

We perform a qualitative analysis on the digital prototype by conducting semi-structured interviews within ING, with particular roles involved in the model. Our approach for the evaluation studies is similar to the stakeholder studies, where we motivate the reasoning behind using a use-case based study and semi-structured interviews.

In Table B.1, we show the evaluation stakeholder study setup, consisting of a mapping to the evaluation stakeholder case study, and stakeholder role. We enlist four stakeholders from varying backgrounds of Model Developer, Model Owner and Domain Expert (who are then given the case study of the Model Owner) to gain insights from different perspective s

For the evaluation study, our high-level approach is as follows :

1. Present a case study where the stakeholder needs to perform a high-level fairness task, such as "ensuring fairness within a particular ML model", or "determining fairness definitions for a particular ML model".



2. Question the steps they would take to accomplish this high-level task (which we will refer to as *fairness steps*)
3. Present the digital prototype to the stakeholder
4. Questions (or observe within the prototype) the *fairness steps* they would take, after navigating through the framework, to accomplish this high-level task
5. Perform a qualitative analysis on what extent the prototype supports stakeholders in F.R.E.M (elicitation and modeling) by observing the effects on the dimensions of **Consideration**, **Specification** and **Communication** that we observe during this study.

The evaluation stakeholder studies are of **semi-structured interview nature** with **aimed duration of one hour (each)**. For each evaluation case study, This case study operates on a dummy **Credit Risk Model** that requires fairness requirements engineering. The evaluation stakeholder case study can be seen in Appendix ??, within the figures of the digital prototype, to get an idea of the type of information exposure presented to the stakeholders.

### Protocol

The evaluation interviews are analyzed in a qualitative manner. Our aim is to observe the interaction of the stakeholder's with the framework and infer insights to see the effects on Consideration, Specification and Communication. To perform the evaluation study, we provide the model developers and model owners with a case study as follows :

### Model Developer

"You are developing a credit-risk model and are in charge of Model Training and Model Evaluation. You have a high-level task of ensuring that the model is fair.

1. **Task 1** : List the steps you would take? This is the status quo procedure for the model developer, for ensuring fairness. With this task, the model developer has a baseline to reflect on their consideration, specification and communication levels after using the prototype.
2. **Task 2** : Review the fairness terminologies shown to you (these consist of fairness definitions, philosophies and metrics extracted from Chapter 2. Indicate which ones you are familiar with? With this task, we get an estimation of how familiar the participant is with fairness notions.
3. **Task 3** : Navigate through the prototype whilst keeping in mind that your task is to ensure the model is fair With this task, we can observe the effects of the different within the framework and discuss improvements. We simulate particular stakeholder roles, and then evaluate on the other stakeholder roles.
  - (a) Reflect : One case study to reflect on, which touches upon proxies
  - (b) Reflect : One dilemma to reflect on, which touches upon the the differing rates of false negatives and false positives between two groups, wherein one group has a protected attribute

- (c) Reflect : Four methods within the machine learning stages with brief description (as seen in Table A.4
- (d) Indicated : The fairness inequalities that are allowed to use, and which are not allowed to use
- (e) Indicated : The fairness definitions are set to mandatory, experiment (means that the definition needs to be explored) and not applicable.
- (f) Mapped : From the fairness definitions, the model developer can review the fairness metric for Binary Classification task. This is the *mapping* between the fairness definition and metric that we find within literature.
- (g) Added : The data-processing stage, activities, fairness steps and trade-offs conducted within these activities is displayed to the model developer. We initially list incomplete fairness steps, and issues within this stage to observe whether model developer's can identify this.
- (h) To do : The model developers can add their own stages, and activities to the model. They can specify fairness steps for each activity.
- (i) Reflect : The model developers are asked if they can for-see trade-offs on dimensions of Table A.3 for the model.

4. **Task 4** : Answer direct questions such as :

- (a) Do you consider more fairness steps than you initially would?
- (b) Do you specify more fairness steps than you initially would?
- (c) Do you consider more fairness trade-offs than you initially would?
- (d) Is the communication on fairness metrics specific?
- (e) Is the communication on fairness methods specific?
- (f) Do the fairness definitions raise consideration?

### Model Owner

"You are in charge of a credit-risk model. You have a high-level task of defining the fairness requirements for the model."

1. **Task 1** : List the steps you would take? This is the status quo procedure for the model developer, for ensuring fairness. With this task, the model developer has a baseline to reflect on their consideration, specification and communication levels after using the prototype.
2. **Task 2** : Review the fairness terminologies shown to you (these consist of fairness definitions, philosophies and metrics extracted from Chapter 2. Indicate which ones you are familiar with?
3. **Task 3** : Navigate through the prototype whilst keeping in mind that your task is to ensure the model is fair. With this task, we can observe the effects of the different within the framework and discuss improvements. We simulate particular stakeholder roles, and then evaluate on the other stakeholder roles.
  - (a) Reflect : One case study to reflect on, which touches upon proxies
  - (b) Reflect : One dilemma to reflect on, which touches upon the the differing rates of false negatives and false positives between two groups, wherein one group has a protected attribute

- (c) Reflect : Fairness Philosophies as mentioned in Table A.2
- (d) To do : Indicate the fairness inequalities that are allowed to use, and which are not allowed to use
- (e) To do : Indicate fairness definitions are set to mandatory, experiment (means that the definition needs to be explored) and not applicable.

4. **Task 4** : Answer direct questions such as :

- (a) Are you considering fairness goals and objectives than more you initially would?
- (b) Are you able to specify fairness goals and objectives better than you initially would?

For Task 3, we guide them to perform primary tasks and repetitive tasks (listed in Table 5.2). Primary tasks are conducted for configuring each level, allowing for the relevant components to be used. Repetitive tasks consist of the same task that have to be conducted for each level, allowing us to see the variations and thought-process invoked as the levels increase.

Monitoring these repetitive tasks allows for inference on how the consideration, specification and communication changes for each stakeholder as new levels are introduced. Since we perform semi-structured interviews, this also gives us an opportunity to delve into further questioning based on the effects of each component and mechanism on the stakeholder (based on their initial answer).

After the stakeholders have performed all primary tasks and repetitive tasks for elicitation and modeling of fairness requirements in the prototype. We ask for a comparison to the status quo procedure of fairness requirements elicitation. The stakeholders are prompted to discuss each component and mechanism for requirements elicitation, in regards to consideration, specification and communication.

## 5.3 Results

### 5.3.1 Analysis

To analyze the evaluation studies, we use the same approach as the stakeholder studies in Chapter 3, wherein we use Atlas.ti, a software that allows for analysis and transcript labeling, to go through the transcripts of the interviews and code relevant occurrences with the following codes:

1. *consideration<sub>satisfier</sub>* : An observation or indication which shows that a particular component(s) of the framework satisfies the consideration challenge.
2. *specification<sub>satisfier</sub>* : An observation or indication which shows that a particular component(s) of the framework satisfies the specification challenge.
3. *communication<sub>satisfier</sub>* : An observation or indication which shows that a particular component(s) of the framework satisfies the communication challenge.
4. *consideration<sub>improvement</sub>* : An observation or indication which shows that a particular component(s) of the framework satisfies the consideration challenge.
5. *specification<sub>improvement</sub>* : An observation or indication which shows that a particular component(s) of the framework satisfies the specification challenge.

Stakeholder	Task Type	Task Description
Model Owner	Repetitive 1	Select relevant fairness definitions for the model
Model Owner	Primary 1	Review case studies
Model Owner	Primary 2	Fill in inequalities levels
Model Owner	Primary 3	Review recommendations
Model Owner	Reference	Review trade-offs
Model Developer	Repetitive 1	Communicate trade-offs
Model Developer	Repetitive 2	Select exploratory mitigation methods
Model Developer	Primary 1	Create model instance
Model Developer	Primary 2	Add model stages
Model Developer	Primary 3	Add model activities
Model Developer	Primary 4	Review fairness definitions
Model Developer	Primary 5	Review recommended fairness metrics
Model Developer	Primary 6	Review possible technical mitigation methods
Model Developer	Reference 1	Review defined social fairness notions

TABLE 5.2: Primary, Reference and Repetitive Tasks

6. *communication<sub>improvement</sub>* : An observation or indication which shows that a particular component(s) of the framework satisfies the communication challenge.

We then review each observation and indication, and derive insights from it.

### 5.3.2 Insights

For each component and mechanism within the framework, we list the observations we find, in terms of Consideration, Specification and Communication. The observations include all stakeholder's perspectives on the components. Based on these insights, we can reflect on Table 4.1, and provide an indication on the impact of the challenge to be addressed, compared to the status quo procedure. Again, we emphasize that the indication should be referred to, in combination with the observations found as the true insights are encapsulating within these observations.

#### 1. Social Fairness Notions

**Observation 1.** : It was difficult for the stakeholders to review the case studies, dilemmas and inequalities and conclude what exactly needs to be done with the information. While the case studies, and dilemmas had "questions to think over", which were aimed at raising consideration, we did not observe this.

**Observation 2.** : Including perspectives is important when reviewing social fairness notions, as model owners/domain experts were more considerate on social fairness notions but preferred to have further discussions on it.

**Observation 3.** : Within industry, translation of social fairness notions to institutional principles is important for model owners/domain experts to grasp the relevance in terms of their institution. It is suggested to add business acumen to further add context to the social fairness notions.

**Observation 4.** For the social fairness notion of inequalities, the exact features mapping to the inequalities can help model developers gain better specification on which features to use and not to use.

**Observation 5.** It is noted that specifying that certain inequalities cannot be used "directly or indirectly", aids in the model developer being more considerate of "indirect influences" compared to the status quo.

**Conclusion :** We find that Consideration is raised to a limited extent, compared to status quo.

#### 2. Fairness Definitions

**Observation 1.** Displaying which fairness definitions are mandatory or intended for experimentation makes it clear for the stakeholder on what "fairness" means for their model.

**Observation 2.** Displaying an intuition behind the fairness definition, that translates it to the ML Model's context can make model developer put the fairness definition into context. For example, one model developer highlights that depending on the model, the outcome that needs to be "fair" can differ. Within a credit-risk model, it can be the probability of a subject being classified as "likely to default" but in another model, for example "collections", the outcome can be related to "".

**Observation 3.** The specification of fairness definitions on what can be used and not, requires discussion and more perspectives to be collected, in order to avoid bias occurring for the personnel setting the fairness definition.

**Conclusion :** We find that Specification and Consideration is raised, compared to status quo.

### 3. Fairness Metrics

**Observation 1.** Displaying the fairness metric for the fairness definition set is beneficial as model developers can identify the main fairness step they need to ensure.

**Conclusion :** We find that Specification is raised, compared to status quo.

### 4. Machine Learning Pipeline Stages :

**Observation 1.** Ability to define stages within the Machine Learning Pipeline can be useful, in comparison to having pre-defined stages. This can allow the stages to be customized to the general stages that a model developing team uses.

**Observation 2.** Model developers lacked evidence for the fairness steps performed within the ML Pipeline. For example, if a fairness step of "Sub-sampling for a particular group" is listed, viewing the distributions can be helpful for conducting further fairness steps in the pipeline. It is also suggested that adding Github Commit Codes to these fairness steps can improve specification, as the developer can now navigate to the code and "review/amend", further addressing specification to the framework. These observations can also be regarded as improvement needed to Communication.

**Observation 3.** We observe that model developer's consider the previous stages occurred within the ML pipeline and review the fairness stages and trade-offs specified within them. One participant was able to spot issues regarding previous fairness steps taken in the ML pipeline.

**Conclusion :** We find that Specification, Consideration and Communication is raised, compared to status quo.

### 5. Machine Learning Pipeline Activities:

**Observation 1.** Model developers specify a few activities within the particular stage, and list fairness steps within for these activities. The act of specifying activities can aid consideration in thinking of fairness steps to conduct for a particular activity. For example, one model developer was able to realize that one more fairness step can be conducted as they specified different activities.

**Observation 2.** Asking for motivations for adding fairness steps within activities prompted consideration of model developers on what other steps they needed to conduct.

**Conclusion :** We find that Specification, Consideration and Communication is raised, compared to status quo.

### 6. Technical Mitigation Methods

**Observation 1:** Model Developers were able to grasp the methods listed within the prototype, indicated that brief method description and names are specific enough. An improvement suggestion was to include links to tools that can help mitigate fairness.

**Observation 2.** By displaying various methods for a particular stage, model developers can consider more methods of conducting a particular fairness step and were considering what the trade-offs could be for using each method.

**Conclusion :** We find that Consideration is raised, compared to status quo.

## 7. Trade-offs

**Observation 1.** Model developers were able to for-see a few trade-offs in the model, but when presented with the trade-off dimensions, model developers were able to for-see trade-offs in more dimensions than initially.

**Observation 2.** Model developers lacked the specification on what was considered a trade-off or not. For example, one model developer eluded that usually a trade-off occurs when e.g. model performance is outside of a particular range. We can infer that some form of specification is needed in terms of the ranges that are considered "unacceptable" and therefore turn into a trade-off.

**Conclusion :** We find that Specification, Consideration and Communication is raised, compared to status quo.





## Chapter 6

# Conclusion

In this Chapter, we reflect on this project, to outline and discuss final outcomes. We started out by formulating four research questions in line with the problem statement of this project. We reflect on each research question, to outline the contribution we make, insights we find, limitations and opportunities for future work.

### 6.1 Reflection

#### 1. Research Question 1

What are the state-of-the-art practices to support fairness requirements elicitation and modeling?

##### **Contribution**

A overview of methods, practices and tasks that can be relevant to consider in eliciting, modeling (including specifying) fairness requirements.

##### **Insights**

We obtain an overview of potential fairness requirements elicitation and modeling practices that can be used stemming from various disciplines. Interestingly, we find that mappings between these disciplines do exist, and can aid in connecting fairness requirements elicitation and modeling.

##### **Limitations**

The limitation of this contribution, is that we do not perform exhaustive studies but rather infer from survey papers to gain an overview.

##### **Opportunities**

More literature studies that focuses on deriving these connections between, for example, fairness philosophies and fairness definitions can really aid in allowing fairness concepts from different disciplines to be reflected upon in fairness in Machine Learning.

#### 2. Research Question 2

What are the institutional challenges when conducting fairness requirements elicitation and modeling?

##### **Contribution**

A study of institutional challenges to address when engineering fairness requirements, specifically eliciting, modeling (including specifying) fairness requirements.

##### **Insights**

By interviewing multiple stakeholders, from various backgrounds, involved in three different types of Machine Learning models, at ING Bank, we are able to gain insights on institutional challenges regarding fairness requirements engineering (specifically, eliciting and modeling).

We find that the challenge within Fairness Requirements Elicitation and Modeling lies with prompting stakeholders to consider aspects of fairness (socially, and technically) that go beyond their current level of knowledge. We also find that the way of communication is important to consider within industry. If a solution for F.R.E.M. is being built, then multiple stakeholders with different backgrounds are likely to be involved, and the solution should take into account these different backgrounds and roles. Lastly, specification is an important challenge to address for F.R.E.M. This is because in order to move from Fairness Requirements Elicitation to specification to modeling, there needs to be empirical and specific information on what needs to be done, or what decisions are made regarding fairness for the ML model.

These three challenges can also be seen as clashing, for example one can ask "When should I stop considering, and start specifying"?. Or one can ask, "Is it not mandatory to go through the theory behind fairness, regardless of the stakeholders' background, to ensure informed decisions are being made". There is a balance required between consideration, specification and communication and of course, industry constraints such as time and resources.

### **Limitations**

The limitation of our stakeholder study, is that there is only one participating institution, namely ING Bank N.V. Hence, we do not consider institutions from different industries, that may have different stakeholders or procedures regarding requirements engineering. This limits us from claiming that these three challenges apply to every institution in the industry.

However, we see that the stakeholders involved in our study, are a model owner (comparative to a product owner), model developer and domain expert. So, whilst our institutions may have additional stakeholders with different backgrounds, we can still assume that when a ML model is being developed, there is someone who is developing it technically, someone who is in charge of the model and the decisions made with it, and there is some knowledge on the domain that the model is being built in. These responsibilities can be mapped to the model owner, model developer and domain expert.

There was limited related work regarding challenges within industry for fairness requirements engineering in literature, we performed a qualitative study to get a sense of what these challenges may look like. We did not perform a set of new interviews or surveys to verify our findings with stakeholders. This can be a limitation for the framework, as the framework is designed towards addressing these challenges.

### **Opportunities**

With this project, an opportunity for future work is on further profiling the stakeholders, to understand the most efficient manner in displaying fairness to these stakeholders, in order to derive requirements for fairness.

## **3. Research Question 3**

What are the requirements and design of the framework?

**Contribution**

A framework aimed at supporting stakeholders involved in Machine Learning model development to elicit, model (including specify) fairness requirements for a ML model whilst addressing institutional challenges discovered in Research Question 2.

**Insights**

We are able to derive a workflow that combines various techniques for deriving fairness goals, and connects it to the technical mitigation methods, and stages involved within the Machine Learning Model Pipeline. We list this as an insight, because this workflow combines different parts of disciplines within Fairness in Machine Learning to present a cohesive system. This directly addresses a lack in literature stated by [34] and [38], the urgent need for internal processes for developing fair ml models (from the start).

**Limitations**

Firstly, there is a lack of verification of the requirements and insights that the framework is built upon. Although we do constantly reflect on literature, the background study and stakeholder study observations to derive the requirements, there is still not a clear verification with the stakeholders on whether the requirements are relevant.

Another limitation is that the requirements can be seen as too generic to re-apply for building a new framework in future research. Here, the evaluation of the framework can bring to light new requirements, such that the framework is can be improved, but this is then specific to the M.L.F.F.

Regarding the framework, the limitations lie within the lack of constraints and behaviors specified for the framework. For example, we do not what actions should absolutely be constrained to prevent accidental misuse of the framework. For example, one can utilize their confusion matrix results, to then derive and justify a fairness definition for the ML model. We also do tackle the human bias added by the fact that humans are still involved with using the framework.

**Opportunities**

Before deriving the framework, we include the requirements and insights that it is based on, which allows these requirements to be re-purposed, or re-evaluated and further improve (or build) a framework/tool.

Regarding the framework, it contains many components (which this project addresses in a more high-level manner) that can be further developed to support the framework and its goals. For example, research on an interactive case-study that results in raising consideration of stakeholders can lead to different (maybe improved) evaluation results and impacts.

**4. Research Question 4**

To what extent does the framework support fairness requirements elicitation and modeling for the institutional challenges identified?

**Contribution**

A qualitative study on the effects of the framework in addressing the institutional challenges in fairness requirements elicitation, and modeling. This includes presenting any additional findings or intriguing observations found

in engagement with the framework, and performing requirements elicitation and modeling.

### Insights

In terms of the framework, we can divide our insights into four categories, namely :

- (a) **ML Pipeline and Technical Mitigation Methods** : We discover that simply visualizing and recording fairness steps (along with motivations) on customized activities and stages can raise consideration on what fairness steps need to be taken. Additionally, this addresses specification in an industry specific way. We find that modeling fairness requirements is an iterative process, wherein a developer may need to perform certain stages again, to compare and evaluate the trade-offs. For example, a developer in charge of model evaluation and optimization may find a trade-off between two groups and ensuring a fairness definition for both groups. One can think, that the steps to take after detecting this trade-off, can be in performing additional sub-sampling tasks, or reviewing the distributions of each group within the data processing stage to decide further actions. Here, the insight we gain is that, facilitating iterative-ness within fairness requirements modeling is necessary and a potential way to do this, is to simply record fairness steps, trade-offs and evidence for the entirety of the ML pipeline of the model. We also find that allowing for customized activities and stages can be impact-ful, in terms of specification and consideration, for the ML model developers.

We also find that by simply showing different technical mitigation methods one can consider for different stages of the ML pipeline, developers can consider more methods to apply. We also find that these methods do not necessarily need to be a specific algorithm, but more high-level methods are also effective in prompting consideration. We find that one reason for this, is because model developers are familiar with high-level technical methods, even though they may not know specific fairness mitigation algorithms and so forth.

- (b) **Social Fairness Notions** : We discover that it is important for the social fairness notions to really be translated into the social context of the model, for stakeholders to make considerations and specifications. We do find that social fairness notions such as fairness philosophies, fairness inequalities can help consideration for the stakeholders, in terms of realizing that discussion is required. It is not enough to enable specification of the fairness requirements.
- (c) **Fairness Definitions, Fairness Metrics, Mappings** : We discover that using mappings to specify fairness definitions, and display the equivalent fairness metric to the model developer allows the model developer to know what fairness means for their model, in a technical manner. It also allows the model developer to infer the main task that needs to be done, to ensure fairness within the ML model.

### Limitations

One limitation for the evaluation is the limited amount of participants, namely four. We try to strive for various backgrounds within these stakeholders to

gain first insights on the framework, we do not perform extensive evaluation with a particular role.

Another limitation of the evaluation, is the semi-structured interviews. Our motivation on performing semi-structured interviews is that we anticipate that the results are subjective and can provide different insights which we may need to delve deeper on. This also means that the interview can focus towards the stakeholder, meaning that we might miss out on other insights that other stakeholders may have provided.

There is also a limitation in terms of the evaluation strategy and time constraints. We have one hour with the stakeholders, we limit the information exposure and stick to basic tasks that need to be done with the framework. This is aligned with the goal of getting first iteration of insights into the framework, it could elude to the first insights not being completely sound, and a validation check may be required.

Additionally, relating to the time constraint, while we access all components to see if consideration, specification and communication is enabled, we do not necessarily evaluate into the negative impacts that the components may have on these three challenges.

Lastly, the framework is not tested for complex models being built. We can expect that within industry, if this framework would need to be applied for a complex ML model (or merely a large project), the requirements for the framework may be different (e.g. more specification is needed than currently found).

### **Opportunities**

With the insights gained from the framework, future work can develop and improve the framework to address the limitations and improvements. For example, more specification can be added within the framework, to include Github commit code references, figures, or statistics. Furthermore, since we build a digital prototype of the framework, one can integrate tools for actually mitigation fairness within algorithms into this prototype to create a solution that allows developers to model the requirements in the framework itself. Another idea could be allowing crowd-sourcing of relevant fairness definitions for the ML model, within the institution via implementing a quiz/form into the digital prototype. In fact, the digital prototype enables a lot of opportunity for expansion in various areas of fairness in machine learning.

## **6.2 Conclusion**

With this project, we provided a base from developing fairness requirements engineering solutions for the industry, to enable Machine Learning models to address fairness at an early stage of development. This field of work is limited, and by provided insights on potential frameworks, challenges and an overview of how a solution for fairness requirements engineering could be developed, we encourage further research and investigation into developing multiple iterations of the framework, that can eventually support fairness requirements engineering.



## Appendix A

# Appendix A

Below, we discuss the traps, the solution and how we understand it.

- **The Framing Trap**

"Failure to model the entire system over which a social criterion, such as fairness, will be enforced"

The main point here touches upon the limitations of an algorithmic frame. An algorithmic frame is defined as an abstraction of the representations (of data) and labeling (outcomes). An evaluation of the algorithmic frame focuses on how the outcome is effected by the inputs, for example, whether the algorithm has good generalization capabilities or whether the algorithm has good accuracy on training data. [63] argues that algorithmic framing is focuses on the improvement of the relationship between the input and outputs, and limited in its capacity to accommodate fairness goals. Another level of frame, is the **data frame**, which then extends the **algorithmic frame**, to incorporate consideration for data representations themselves, and what they mean in terms of quality. This **data frame**, already allows for incorporating certain fairness goals. For example, [Feldmen et al] shows that removing bias in training data before it is actually passed on to the training for the model, can improve fairness. [63] states that even this **data frame**, is mostly geared towards mathematical implementations of fairness without consideration for contextual understanding. Therefore, the **socio-technical** frame can be considered, which recognizes that the machine learning model is part of the socio-technical system.

The recommendation provided for this trap is "is heterogeneously framed so as to include the data and social actors relevant to the localized question of fairness". Here, heterogeneously entails the the addition of people, social systems, institutional environments, regulatory systems and different technical parts of the model to be considered simultaneously [John Law].

Inequality	Example
Natural inequality	Disability at birth
Socioeconomic inequality	Parents'/guardians' assets
Talent inequality	Intelligence, skills, employment prospects
Preference inequality	Saving behavior, cultural prioritization of values associated with economic opportunities
Treatment inequality	Discrimination in job market and education system affecting income stability

TABLE A.1: Levels of inequalities [46]

Philosophical perspective	Acceptable inequalities	Unacceptable inequalities
Formal equality of opportunity / procedural fairness (Greenberg, 1987)	Any inequality as long as the opportunity was open to all	Treatment inequality
“Fair equality of opportunity” (Rawls, 1999, 2001)	Natural, talent, and preference inequalities	Socioeconomic, treatment inequalities
Rawlsian EOP + Difference principle (Rawls, 1999)	Natural, talent, and preference inequalities, plus any inequality benefiting the most disadvantaged society members in long-term impact	Socioeconomic, treatment inequalities
Equality of outcome / condition / welfare (Greenberg, 1987)	None - all members should get the exact same outcome	All
Luck egalitarianism (Dworkin, 1981)	Effort-based inequalities (e.g. preference)	Circumstances (e.g. natural inequality)
Equality of freedom / autonomy (Sen, 1992)	Inequality resulting in “genuinely free” choices	Any inequality hindering freedom
Sufficiency / Equality of capability (Walzer, 1983)	Any inequality as long as everyone is above the level of sufficiency	Any resulting in people falling below sufficiency levels
Prioritarianism (Scheffler, 1994; Parfit, 1991)	Any inequality reduction should prioritise resource allocation to those who are worst off	None as long as the worst off are prioritised
Desert (Kagan, 1999, 2014)	Any inequality based on what he/she “deserves”	Any inequality that does not equate to the person’s deserving

TABLE A.2: Fairness Philosophies and corresponding inequalities mentioned in [46]

Fairness Dimension	Specification	Description
Procedural Fairness	Process Control	The ML model shall provide the individual sufficient control over the procedure.
	Decision control	The ML model shall provide the individual sufficient influence over the decision outcome.
	Consistency Across individuals	The ML model shall apply decision-making procedures consistently across individuals.
	Across time	The ML model shall apply decision-making procedures consistently across time.
	Impartiality	The ML model shall be neutral and guard against those with an interest in the decision.
	Bias suppression	The ML model shall suppress undesirable outcome biases.
	Internal data quality	The ML model shall use data that sufficiently represents the real world.
	External data quality	The ML model shall use data that is sufficiently usable and valuable for its intended functionality
	Model performance	The ML model shall use predictive models with high performance.
	Review	The ML model shall easily allow individuals to review the relevant information leading up to the decision and the accountable entity.
	Contest	The ML model shall easily allow individuals to contest adverse or incorrect decisions.
	Human oversight and correction	The ML model shall allow human operators to oversee decision-making and correct decisions.
	Representative subgroup involvement	All phases of the ML model shall involve the important subgroups in the population of individuals affected by the ML model.
Target representation	The ML model shall aim for a certain representation of subgroups in the target population.	
Lawfulness	The ML model shall operate in accordance with applicable laws and regulations.	
Justification	The goal and practices of the ML model shall be justifiable within the organizational and societal values.	
Distributive Fairness	Distributive Norms	The ML model shall allocate the resources (outcomes) in a manner consistent with its goals.
	Characteristics	The ML model shall take specific relevant characteristics into account.

TABLE A.3: Fairness Dimensions (Procedural and Distributive) specified by [45]



Techniques Stage	Description
Blinding	Methods aim to make the classifier immune to one or more sensitive variables, wherein there is no outcome differentiation based on a particular sensitive variable.
Casual Methods	Methods aim to uncover casual relationships by finding underlying dependencies within the data (used in training the ML Model)
Sampling	Methods aim to correct training data by removing bias.
Subgroup Analysis	Methods can aim to identify sub-samples (groups) within the training data that are disadvantaged by the classifier to evaluate the model.
Transformation	Methods that learn new representations of data (for example, a mapping or projection function) to ensure fairness.
Relabeling	Methods flip/modify the dependent variables, to evaluate the outcome changes that occur for different groups.
Perturbation	Methods aim to change the distribution of one or more variable in the training data, as a way to <b>repair</b> the data.
Reweighting	Methods aim to assign weights to particular instances of training data (without changing the data itself).
Regularization	Methods aim to penalize the classifier for discriminatory practices, by adding one or more penalty terms.
Constraint Optimization	Methods aim to constrain the classifier loss function operating on the confusion matrix
Adversarial Learning	Methods aim to use an adversary to try to determine whether a training model is robust enough. Methods can penalize a model if the sensitive variable is predictable from the dependent variable.
Calibration	Methods aim to ensure that the proportion of positive predictions is equal to the proportion of positive examples for all subgroups.
Thresholding	Methods aim to find regions of the posterior probability distribution of a classifier where favored and protected groups are both positively and negatively classified, and determine threshold values.

TABLE A.4: Fairness Techniques (for Binary Classification) mentioned in [15]

- **The Portability Trap** "Failure to understand how re-purposing algorithmic solutions designed for one social context may be misleading, inaccurate, or otherwise do harm when applied to a different context"

Here, it is argued that division by task of the algorithm can lead to misconception that solutions can be used without contextual consideration. From our background study, we identified that Machine Learning Tasks can be classification, regression, re-enforcement learning. The portability trap eludes to saying, that, for example, if classification models are built for different purposes (credit loan decisions, hiring decisions or perhaps which institutional division a consumer's question fall into), then just because the model type is the same, does not mean that the fairness and technical solutions are the same as well. That is, then, dependent on the context. Intuitively, this is evident as well, as classifying a consumer's question into the division can have more leeway for incorrect classifications compared to credit loan decisions (as the impacts of the latter are dire).

The recommendation proposed for this trap is "has appropriately modeled the social and technical requirements of the actual context in which it will be deployed"

From this, we understand that we need to extend beyond types of model tasks to prompt attention to context, and thus prevent the portability trap.

- **The Formalism Trap**

"Failure to account for the full meaning of social concepts such as fairness, which can be procedural, contextual, and contestable, and cannot be resolved through mathematical formalisms"

The trap to fall in here, is to not involve aspects such as procedural fairness, contextual information and contest-ability to supplement mathematical definitions of fairness.

The recommendation for this trap is as follows : "can appropriately handle robust understandings of social requirements such as fairness, including the need for procedurality, contextuality, and contestability (Formalism);"

To this end, we understand that social fairness notions, fairness definitions and fairness metrics need to be considered within the technical system.

- **The Ripple Effect Trap**

"Failure to understand how the insertion of technology into an existing social system changes the behaviors and embedded values of the pre-existing system"

The ripple effect trap addresses that an understanding of intended and unintended consequences for inserting technical systems in pre-existing systems should be made. The socio-technical recommendation for this is : "affects the social context in a predictable way such that the problem that the technology solves remains unchanged after its introduction". Here, we are prompted to take into consideration **what-if** scenarios that can occur to re-interpret the

- **The Solutionism Trap** "Failure to recognize the possibility that the best solution to a problem may not involve technology"

This trap refers to the limitation that technology might add to consider political nuances and fairness solutions, as computationally or even observationally it is too

complex to implement within an algorithm or some form of a simulation system. The socio-technical perspective on this is as follows : "is appropriate to the situation in the first place, which requires a nuanced understanding of the relevant social context and its politics".



## Appendix B

# Appendix B

## B.1 Stakeholder Study A

### B.1.1 Model Owner

Disclaimer

- Consider the scenario as ground-truth for the purpose of this interview (i.e. contradictions to the scenario itself occurring do not count as a valid answer)
- Assumptions : Only model developers, model users and model owners (i.e. validation team has no responsibility) Processes do not occur like a standard bank as you know it. Scenario is 'deliberately detached' from current workings of a bank.
- You are the model owner for this scenario and bank.

Imagine Dina, the loan manager who has worked for more than 15 years at the bank. Imagine Amrani, Dina's long term client, who wants to renew his mortgage loan. Amrani gets a negative response to his application. Amrani calls Dina and is upset about this. Dina reviews Amrani's financials and concludes that Amrani should have been given the loan Dina notices that Amrani is a non-native speaker and has been through some employment issues in the past. Dina decides to reach out to the model developing team.

1. Who do you think Dina's majority discussion will be with and why?
  - (a) You
  - (b) The technical team that developed the model
2. Dina states that there are many complex cases within the bank such as Amrani's that can be considered outliers. She asks you : "How would you communicate the accommodation for these different complexities to your developers?"
3. Then Dina says the following : "I always considered that via these models, if the majority are given the best solution, then it is a win. With these models, how can we make sure that this majority is maximized?" What would be your approach to do this?
4. Which one of the following statements would you most identify with?
  - (a) The data we have has been collected over many years so the models are trained to include multiple status quos that have been present over the years.

- (b) The models have a high predictive accuracy so almost everyone is getting a correct evaluation.
  - (c) We make sure that the data is of high quality.
  - (d) We want to 'replace' human decisions so using the past data is the best approach to take.
5. As a bank, which situation do you consider more?
- (a) Of those to whom I granted a loan, how many will actually not pay?
  - (b) Of those that I decided to reject, how many would actually pay?

### B.1.2 Model Developer

- Consider the scenario as ground-truth for the purpose of this interview (i.e. contradictions to the scenario itself occurring do not count as a valid answer)
- Assumptions : Only model developers, model users and model owners (i.e. validation team has no responsibility) Processes do not occur like a standard bank as you know it. Scenario is 'deliberately detached' from current workings of a bank.
- You are the model developer for this scenario and bank.

Imagine Dina, who has requested for this model to be built and has worked for more than 15 years at the bank.

1. While discussing with you, Dina tells you : "I really want to put an emphasis on the model being as fair as possible. Could you add this requirement to the overall model requirements? " What types of further questions would you ask her?
2. You proceed with your team (you and your colleague) to develop this model. This requirement of 'fairness', which parts of the ML pipeline would you associate this with?
3. So, before the algorithmic model development, you and your colleague look into the data. Your colleague messages you saying the following : "Hi, after looking at the data, I noticed that there were some missing values. I just removed them. Don't worry, they were less than 1 attributes. Should we get the data approved by the DPO?"
  - (a) Yes, that sounds good.
  - (b) No
4. After creating the model, you do some statistical tests to check dependencies between significance attributes and the 'other' attributes to check correlation. How reliable are these statistically tests? Do you consider nonlinear dependencies?
5. You and your colleague then successfully create a model. You have a few parameters that you can tune. You notice that each hyper-parameter changes the confusion matrix. How do you decide the hyper-parameter values?  
Which situation would you consider more?

- (a) Of those to whom I granted a loan, how many will actually not pay?
  - (b) Of those that I decided to reject, how many would actually pay?
6. News Article : "Men coming from a privileged background have a higher credibility rate than women with the same backgrounds" To what extent would you explore this (in terms of actions)? Would you consider this your responsibility? Who else would you deem accountable in such situations?
7. Out of these four statements, which ones would you tell Dina to assure her that the model is fair?
- (a) The process in which the model assesses is fair.
  - (b) The outcome of the model is fair.
  - (c) We consider our client's situation and try to provide the best solution for them.
  - (d) We try to provide all our clients with enough feedback on why a certain decision was made.
8. Which one of the following statements would you use to communicate how the model was created?
- (a) We have collected over many years so the models should be trained to include multiple status quos that have been present over the years.
  - (b) The model needs to have a high predictive accuracy so almost everyone is getting a correct evaluation.
  - (c) We use data that is of high quality.
  - (d) We want to 'replace' human decisions so we use past data for best results
9. In an automated process, there is a group of people that are given a wrong judgement. Which statement do you think is most appropriate here?
- (a) This is just random error and the state of technology.
  - (b) We need to make sure that we do not make this mistake next time.
  - (c) The more data we have, the better the automated process will become. We just have to wait.

### B.1.3 Domain Expert

#### Disclaimer

- Consider the scenario as ground-truth for the purpose of this interview (i.e. contradictions to the scenario itself occurring do not count as a valid answer)
- Assumptions :
- Only model developers, model users and model owners (i.e. validation team has no responsibility)
- Processes do not occur like a standard bank as you know it. Scenario is 'deliberately detached' from current workings of a bank.
- You are the model user for this scenario and bank

Imagine Dina, the manager of the bank, has worked for more than 15 years at the bank. Imagine Nadia, Dina's long term client who wants a mortgage loan but received a NEGATIVE assessment.

While discussing with you, Dina tells you that Nadia claims to be subject to the gender pay gap in her previous company. I.e., she was receiving less money than her fellow male colleagues. In the past year, she moved into a new job with a higher income but her period of her job is not long enough for a positive loan assessment.

1. Select the response you identify with the most :
  - (a) The model is objective therefore the decision will remain.
  - (b) This is an isolated incident, and we will simply overwrite the decision and make an exception due to long-term relation we have with Nadia.
  - (c) We are truly sorry but this is not our problem.
  - (d) We will talk to the model development team about this incident.
2. Out of these four statements, which ones would you advise Dina to tell Nadia to assure her?
  - (a) The process is identical in which the model assesses is fair.
  - (b) The outcome of the model is fair.
  - (c) We consider our client's situation and try to provide the best solution for them.
  - (d) We try to provide all our clients with enough feedback on why a certain decision was made.
3. Which one of the following statements would you use to communicate what you expect from the model ?
  - (a) We have collected over many years so the models should be trained to include multiple status quos that have been present over the years.
  - (b) The models need to have a high predictive accuracy so almost everyone is getting a correct evaluation.
  - (c) We need to use data that is of high quality.
  - (d) We want to 'replace' human decisions so using the past data is the best approach to take.
4. As a bank, which situation do you consider more?
  - (a) Of those to whom I granted a loan, how many will actually not pay?
  - (b) Of those that I decided to reject, how many would actually pay?
5. In a automated process, there is a group of people that are given a wrong judgement. Which statement do you think is most appropriate here?
  - (a) This is just random error and the state of technology.
  - (b) We need to make sure that we do not make this mistake next time.
  - (c) The more data we have, the better the automated process will become. We just have to wait.



6. Now, suppose that this (misclassified) group of people mostly consist of ‘unprivileged’ groups. But there is not one single ‘unprivileged group’ that stands out so the automated process seems to be fair.

What do you think that the bank’s approach to this should be?

- (a) We have a responsibility to the society and our clients so we need to consider this.
- (b) We rely on statistics, data and technology to make decisions. There is no evidence here of unfairness.
- (c) We cannot check every single case as it is just out of scope and infeasible. We try our best to improve predictive accuracy.

## B.2 Stakeholder Study B

### B.2.1 Model Owner

- Consider the scenario as ground-truth for the purpose of this interview (i.e. contradictions to the scenario itself occurring do not count as a valid answer)
- Assumptions : Only model developers, model users and model owners (i.e. validation team has no responsibility)

Processes do not occur like a standard bank as you know it. Scenario is ‘deliberately detached’ from current workings of a bank.

- You are the model owner for this scenario and bank

Imagine Dina, working at the Payment Collection Service of the bank. Imagine Nadia, who has received multiple payment requests from the bank regarding her vault at the bank. Nadia is irritated as she tried to speak into the telephone multiple times but every time either it did not understand her or it re-directs to the wrong service desk. Nadia was in a quiet room. For this case, let us assume that the person at the service desk is unable to re-direct to another service desk.

1. What could be the cause of this?
  - (a) Nadia did not speak clearly
  - (b) The model sometimes makes mistakes. It is not 100
  - (c) The model may not have been able to understand Nadia.
  - (d) If Nadia tries enough times, eventually it will work.
2. Select the response that you identify with the most:
  - (a) This was an isolated incident but the important thing is that our customers’ information was not revealed to any third parties.
  - (b) This was not a isolated incident but the important thing is that our customers’ information is not revealed to any third parties.
  - (c) The impact on the customer affects our service quality so we need to put in more resources to accommodate all types of clients.
  - (d) There is a potential reputational risk associated with this incident so we need to look into this further.

3. How would you classify the incident that happened with Nadia? Just an estimation...
- (a) Bias
  - (b) Error
  - (c) Under-training of the model
  - (d) Lack of data

How do you see the impact of this in other models in the pipeline?

4. You talk to your developers. After a conversation, you realise that there is a trade-off between 'removing private information accurately' versus 'retaining information to re-direct the customer accurately'

How do you convey the preference of this trade-off to the model developers? Is there a way you can formalize this?

What do you expect the model developers to show to you to prove that they implemented your thought process correctly?

5. Throughout time, it is realised that the accents of non-native speakers are not properly recognised so these people are always re-directed incorrectly.

What do you think will be the viable solution?

- (a) The problem is in the data. Unless that is solved, we cannot do anything.
- (b) The safety (masking/identifying PII) will be worse if we try to retain more information. There is a choice to be made.
- (c) We will need to look into the pipeline to see if and how we can do this.
- (d) We can solve this problem in a short-period of time as we are aware of where the problem occurs.

6. Out of these four statements, which ones would you advise Dina to tell Nadia to assure her about this automated telephone system.

- (a) We understand that you are upset. The technology is new and as we get more data, the technology will improve.
- (b) We will look into accommodating you with a dedicated/alternative solution.
- (c) We will treat this as a mistake we need to improve upon and will talk to the model developers.
- (d) We will look into accommodating you with a direct diversion to a worker at the bank.

## B.2.2 Model Developer

- Consider the scenario as ground-truth for the purpose of this interview (i.e. contradictions to the scenario itself occurring do not count as a valid answer)
- Assumptions : Only model developers, model users and model owners (i.e. validation team has no responsibility)
- Processes do not occur like a standard bank as you know it. Scenario is 'deliberately detached' from current workings of a bank.

- You are the model owner for this scenario and bank

Imagine Dina, working at the Payment Collection Service of the bank. Imagine Nadia, who has received multiple payment requests from the bank regarding her vault at the bank. Nadia is irritated as she tried to speak into the telephone multiple times but every time either it did not understand her or it re-directs to the wrong service desk. Nadia was in a quiet room. For this case, let us assume that the person at the service desk is unable to re-direct to another service desk.

1. What could be the cause of this?
  - (a) Nadia did not speak clearly
  - (b) The model sometimes makes mistakes. It is not 100
  - (c) The model may not have been able to understand Nadia.
  - (d) If Nadia tries enough times, eventually it will work.
2. The model owner comes to you to communicate the following. He says that 'We should make sure that the model is fair towards all customers'. What questions do you ask him to meet his requirement? What are ways you can show him that you have met his requirements?
3. How would you classify the incident that happened with Nadia? Just an estimation...
  - (a) Bias
  - (b) Error
  - (c) Under-training of the model
  - (d) Lack of data

How do you see the impact of this in other models in the pipeline?

4. Throughout time, it is realised that the accents of non-native speakers are not properly recognised so these people are always re-directed incorrectly. The model owner comes to you to ask to solve this problem in a short period of time. Select the statement you identify the most with :
  - (a) The problem is in the data. Unless that is solved, we cannot do anything.
  - (b) The safety (masking/identifying PII) will be worse if we try to retain more information. There is a choice to be made.
  - (c) We will need to look into the pipeline to see if and how we can do this.
  - (d) We can solve this problem in a short-period of time as we are aware of where the problem occurs.
5. You talk to your model owner. After a conversation, you state that there is a trade-off between 'masking private information' versus 'retaining information to re-direct the customer accurately' What do you do as a model developer? Have you encountered this type of scenarios?
6. After a while, the 'original model' is being used in a recommendation system for predicting customer chats within a chatbot. You talk to the people and they say the following 'Because the model anonymizes PII, the recommendation system will not learn on personal information so it will not discriminate'. What is your thoughts on this statement? Based on this, what actions would you take?

### B.2.3 Domain Expert

- Consider the scenario as ground-truth for the purpose of this interview (i.e. contradictions to the scenario itself occurring do not count as a valid answer)
- Assumptions : Only model developers, model users and model owners (i.e. validation team has no responsibility)

Processes do not occur like a standard bank as you know it. Scenario is 'deliberately detached' from current workings of a bank.

- You are the model owner for this scenario and bank

Imagine Dina, working at the Payment Collection Service of the bank. Imagine Nadia, who has received multiple payment requests from the bank regarding her vault at the bank. Nadia is irritated as she tried to speak into the telephone multiple times but every time either it did not understand her or it re-directs to the wrong service desk. Nadia was in a quiet room. For this case, let us assume that the person at the service desk is unable to re-direct to another service desk.

1. What could be the cause of this?
  - (a) Nadia did not speak clearly
  - (b) The model sometimes makes mistakes. It is not 100
  - (c) The model may not have been able to understand Nadia.
  - (d) If Nadia tries enough times, eventually it will work.
2. Select the response that you identify with the most:
  - (a) This was an isolated incident but the important thing is that our customers' information was not revealed to any third parties.
  - (b) This was not a isolated incident but the important thing is that our customers' information is not revealed to any third parties.
  - (c) The impact on the customer affects our service quality so we need to put in more resources to accommodate all types of clients.
  - (d) There is a potential reputational risk associated with this incident so we need to look into this further.
3. Out of these four statements, which ones would you advise Dina to tell Nadia to assure her about this automated telephone system.
  - (a) We understand that you are upset. The technology is new and as we get more data, the technology will improve.
  - (b) We will look into accommodating you with a dedicated/alternative solution.
  - (c) We will treat this as a mistake we need to improve upon and will talk to the model developers.
  - (d) We will look into accommodating you with a direct diversion to a worker at the bank.
4. How would you classify the incident that happened with Nadia? Just an estimation...

- 
- (a) Bias
  - (b) Error
  - (c) Under-training of the model
  - (d) Lack of data
5. You talk to your developers. After a conversation, you realise that there is a trade-off between 'removing private information accurately' versus 'retaining information to re-direct the customer accurately' How do you convey the preference of this trade-off to the model developers? Is there a way you can formalize this? What do you expect the model developers to show to you to prove that they implemented your thought process correctly?
6. Another complaint comes your way. Nadia and many customers claim that they feel uncomfortable with revealing their information to a machine. How do you envision this problem being addressed?
- (a) Provide them with evidence that the model is accurate enough to remove sensitive information.
  - (b) Assure them that the model is constantly being improved.
  - (c) Pull back the model and re-release it with an even higher accuracy than before.
  - (d) Automation is the future. The customers will have to accept our decisions.

Model Type	Roles	Question	Objective Mapping
Binary Classification	Model Owner	A.1.1. Question 1	Objective 2b
Binary Classification	Model Owner	A.1.1. Question 2	Objective 2a
Binary Classification	Model Owner	A.1.1. Question 3	Objective 1a, 1b
Binary Classification	Model Owner	A.1.1. Question 4	Objective 1a, 1b
Binary Classification	Model Owner	A.1.1. Question 5	Objective 2c
Binary Classification	Model Developer	A.1.2. Question 1	Objective 2a
Binary Classification	Model Developer	A.1.2. Question 2	Objective 1a
Binary Classification	Model Developer	A.1.2. Question 3	Objective 1b
Binary Classification	Model Developer	A.1.2. Question 4	Objective 1a, 1b
Binary Classification	Model Developer	A.1.2. Question 5	Objective 2b, 2c
Binary Classification	Model Developer	A.1.2. Question 6	Objective Objective 2b
Binary Classification	Model Developer	A.1.2. Question 7	Objective 2a
Binary Classification	Model Developer	A.1.2. Question 8	Objective 1a, 1b
Binary Classification	Model Developer	A.1.2. Question 9	Objective 1a, 1b
Binary Classification	Domain Expert	A.1.3. Question 1	Objective 2a, 2b
Binary Classification	Domain Expert	A.1.3. Question 2	Objective 1b
Binary Classification	Domain Expert	A.1.3. Question 3	Objective 1a, 1b
Binary Classification	Domain Expert	A.1.3. Question 4	Objective 2c
Binary Classification	Domain Expert	A.1.3. Question 5	Objective 1a, 1b
Binary Classification	Domain Expert	A.1.3. Question 6	Objective 1b, 2c, 2a
NLP, Entity Detection	Model Owner	A.2.1. Question 1	Objective 1a, 1b
NLP, Entity Detection	Model Owner	A.2.1. Question 2	Objective 1b, 2c
NLP, Entity Detection	Model Owner	A.2.1. Question 3	Objective 1a, 1b
NLP, Entity Detection	Model Owner	A.2.1. Question 4	Objective 2a, 2b
NLP, Entity Detection	Model Owner	A.2.1. Question 5	Objective 1a, 1b
NLP, Entity Detection	Model Owner	A.2.1. Question 6	Objective 2c
NLP, Entity Detection	Model Developer	A.2.1. Question 1	Objective 1a, 1b
NLP, Entity Detection	Model Developer	A.2.2. Question 2	Objective 2a, 2b
NLP, Entity Detection	Model Developer	A.2.2. Question 3	Objective 1a, 1b
NLP, Entity Detection	Model Developer	A.2.2. Question 4	Objective 1a
NLP, Entity Detection	Model Developer	A.2.2. Question 5	Objective 2b, 2c
NLP, Entity Detection	Model Developer	A.2.2. Question 6	Objective 1a, 1b, 2a, 2b
NLP, Entity Detection	Domain Expert	A.2.3. Question 1	Objective 1a, 1b
NLP, Entity Detection	Domain Expert	A.2.3. Question 2	Objective 1b, 2c
NLP, Entity Detection	Domain Expert	A.2.3. Question 3	Objective 2a, 2b
NLP, Entity Detection	Domain Expert	A.2.3. Question 4	Objective 1a, 1b, 2c

TABLE B.1: Stakeholder Study Mapping to Objectives

## Appendix C

# Appendix C

### C.1 Evaluation Stakeholder Study in Digital Prototype

In this section, we show that evaluation stakeholder study within the digital prototype used for the evaluation tasks to be conducted. We also map the Figures to the corresponding component within the M.L.F.F.

#### C.1.1 Introduction and Common Components

In Figure C.2 - Figure C.5, we show the introduction and common components visited by both Model Owner and Model Developer.

#### C.1.2 Model Developer

In Figure C.6: Figure ??, we show the components visited by the model developer.

#### C.1.3 Model Owner

In Figure C.15: Figure C.17, we show the components visited by the model developer.

### ✔ Why are you here?

This tool is aimed towards fairness requirements engineering for machine learning models.

So, what makes fairness requirements engineering different from normal requirements engineering?

Well, fairness requirements can be :

- **Subjective**
  - E.g. Different people may have different perspectives on what fairness means, even within the same organization
- **Uncertain**
  - E.g. As time goes on, new requirements may come to light, or simply may not be realizable as of this moment
- **Variable**
  - E.g. For different countries, or different groups of people, different fairness requirements are applicable

This tool will provide you will tasks that can help you elicit and model fairness requirements.

- **Elicitation**
  - Means you can explore fairness in the context of your model, and formally define it.
- **Modeling**
  - Means you can discuss the trade-offs, realizability and further investigate any actions to take regarding the formally defined fairness requirements.

FIGURE C.1: Introduction Text

### Choose your model

If your model name is not in the list, please add a new model

Model

Add new model

Enter the model name

Enter model description

FIGURE C.2: Component : Model

### 🔗 Enter your role

Select your role

Model Owner

A few questions first...

FIGURE C.3: Selecting Role



## Case Studies

Select model domain

Commercial lending operations

Show relevant case studies

Case Study 1 : More disabled candidates recieved a negative qualification on a loan decision model, due to socio-economic level. Upon re-assessment, many candidates were actually eligible for the loan. The model had been using location as a proxy for inferring disability within the candidate and many more disabled candidates got a false negative decision, compared to the average rate of getting a false negative decision.

FIGURE C.4: Component : Case Studies

## Dilemmas

See common causes of bias

See social context

Nina, is applying for a loan in Country X, and submits her data. The ML model does not use the attribute of gender in its calculations, but was able to figure that Nina's gender by the type of car she owned (a.k.a. proxy). Based on the computation of the ML model, Nina has a higher chance of being rejected and a feature that plays a part in this, in the type of car she owns.

### Questions to think over

1. Does further evaluation need to be done on this, to detect discrimination of gender?
2. How representative is the data to the current economic situation? Could there be historical bias?
3. Should this have been controlled before the model was deployed?

FIGURE C.5: Component : Dilemmas

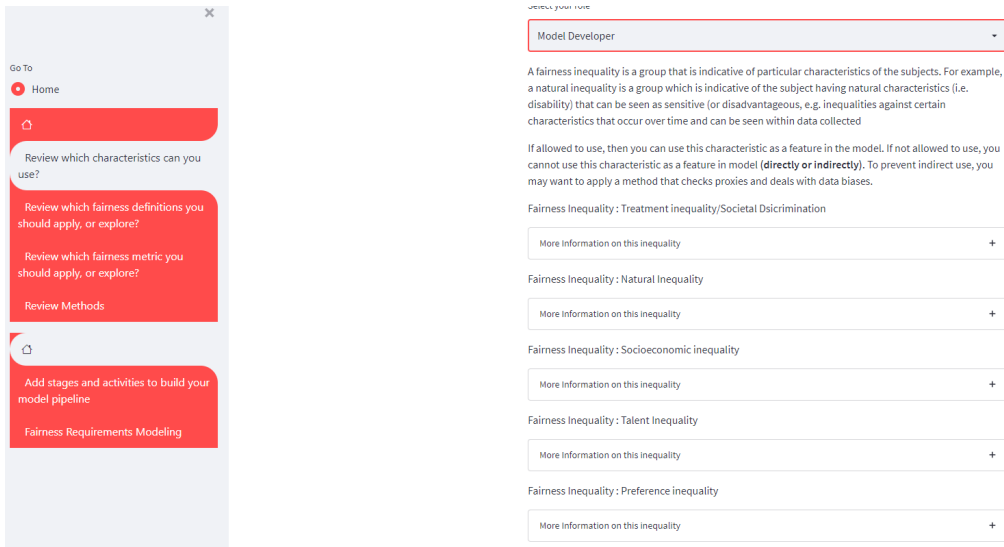


FIGURE C.6: Component : Social Fairness Notions - Reviewing inequalities

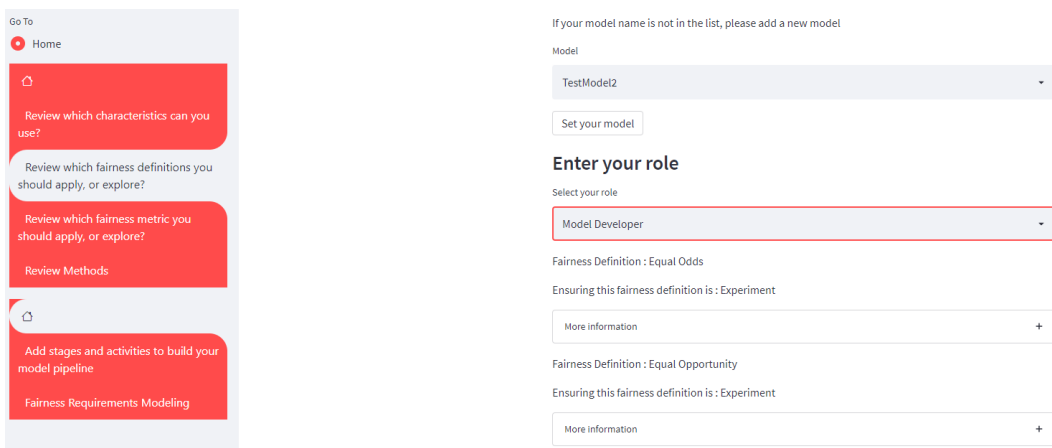


FIGURE C.7: Component : Fairness Definitions - Reviewing fairness definitions

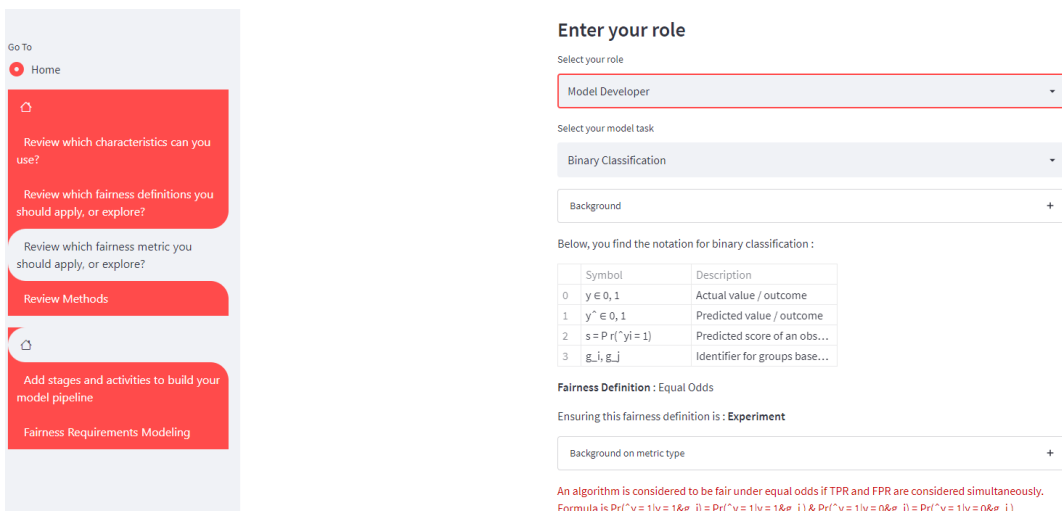


FIGURE C.8: Component : Fairness Metrics

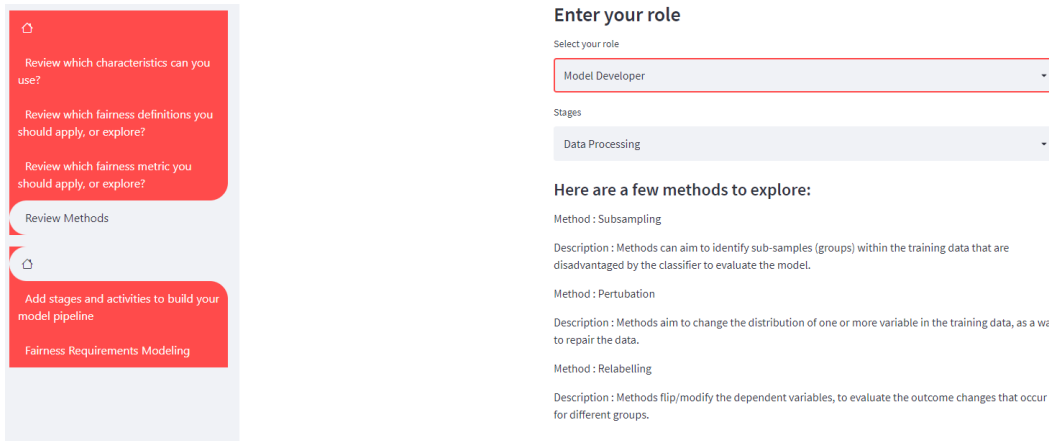


FIGURE C.9: Component : Technical Mitigation Methods

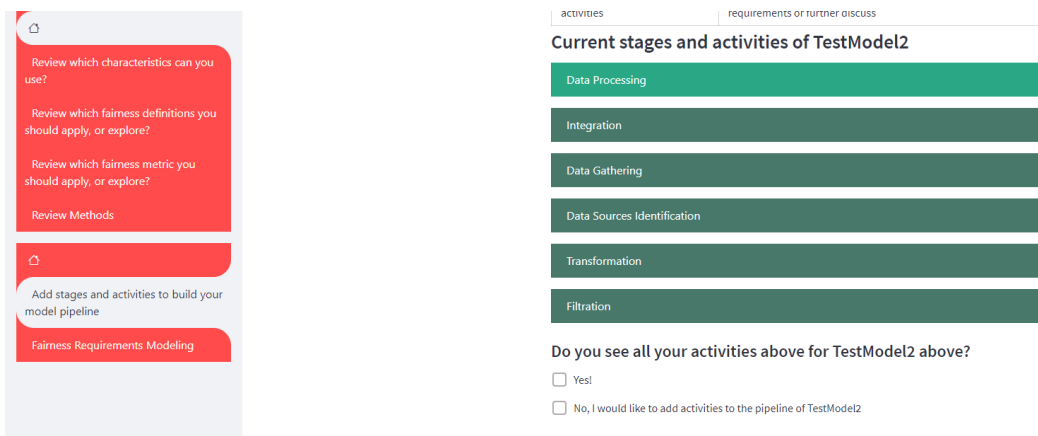


FIGURE C.10: Component : ML Pipeline Stages

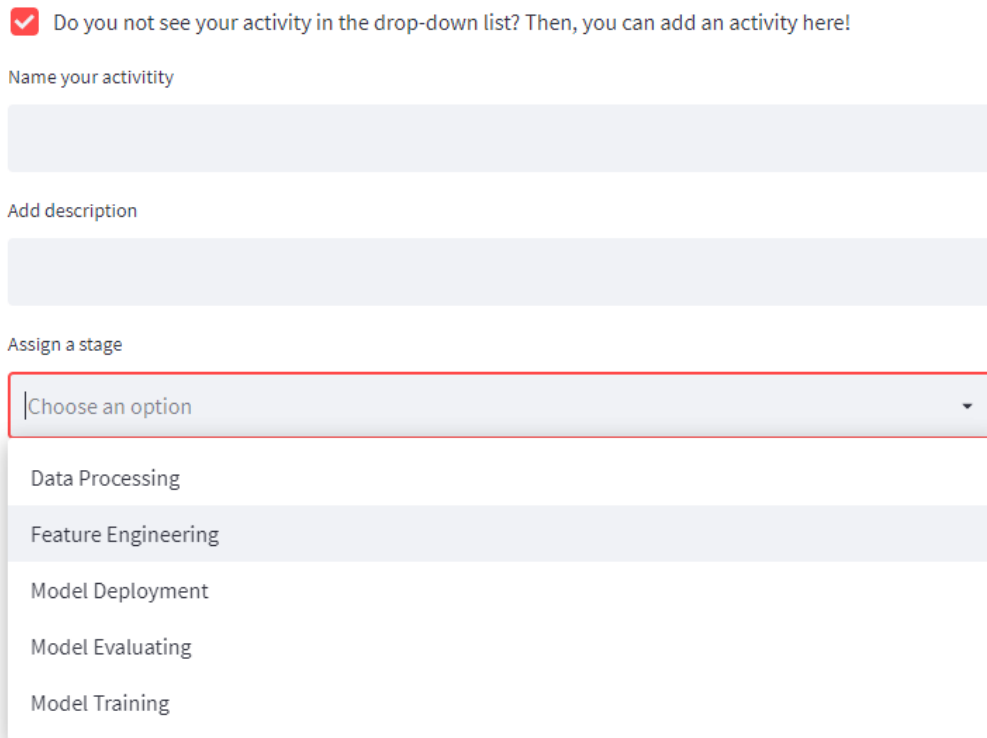


FIGURE C.11: Component : ML Pipeline Activities - Adding an activity to a stage

View other stages of the pipeline to see the actions taken

- Stage : Data Processing
  - Activity : Data Gathering
    - Fairness Steps taken : None
    - Trade-offs : Internal Data Quality may involve historical bias from past policies of assigning a credit-risk scoring
  - Activity : Data Transformation
    - Fairness Steps taken : Perturbation, i.e. distribution of data changed on race, gender
  - Activity : Integration
    - Fairness Steps taken :
      - Missing values removed for race, gender
      - Feature of race, gender removed

Record fairness steps for your activities

Record trade-offs on the following dimensions

FIGURE C.12: Component : ML Pipeline Stages and Activities - Reviewing the ML Pipeline

Add your fairness steps!

Name your fairness step

Add description

Assign a stage

Data Processing ▼

Assign an activity

Compression ▼

FIGURE C.13: Component : ML Pipeline Stages and Activities - Adding a fairness step to the ML Pipeline

Record trade-offs on the following dimensions

The ML model shall use data that sufficiently represents the real world.

Enter the trade-offs on Internal Data Quality

The ML model shall provide the individual sufficient control over the procedure.

Enter the trade-offs on Process Control

The ML model shall suppress undesirable outcome biases.

Enter the trade-offs on Bias Suppression

FIGURE C.14: Component : Trade-offs

Go To

- Home

View previous fairness case studies recorded in the model domain

View dilemmas set by your institution for this model

Indicate which characteristics should not be discriminated against

Indicate Fairness Definitions to display to the model developer

Overview Fairness Requirements Specification

**Natural Inequality**

- Description : Natural inequality stems from differences in age, health or other physical characteristics,
- Intuition : Characteristics such as Disability at birth

Select whether Natural Inequality applicable or not

Allowed to use  Not allowed to use

Please state your motivation

**Preference inequality**

- Description : Cultural prioritisation of values associated with economic opportunities,
- Intuition : Characteristics such as Saving behaviour

Select whether Preference inequality applicable or not

Allowed to use  Not allowed to use

Please state your motivation

FIGURE C.15: Component : Social Fairness Notions - Specifying inequalities

✓ Based on your selection, you can get guidance on fairness philosophies that are relevant

## Guidance for selecting fairness philosophies based on your selected inequalities

**Equality of Freedom** can be relevant to consider as it allows for the following inequalities to be considered during the model ("Equality of autonomy is a political philosophy concept that argues "that the ability and means to choose our life course should be spread as equally as possible across society"— i.e., an equal chance at autonomy or empowerment. Equality of autonomy strives to spread empowerment widely so that "given their circumstances", people have more "choice and control" that considers these inequalities as acceptable : natural\_inequality, talent\_equality, preference\_inequality, socioeconomic\_inequality, treatment\_inequality and that considers these inequalities as unacceptable : ; ")

**Fair Equality of Opportunity** can be relevant to consider as it allows for the following inequalities to be considered during the model ("Fair Equality of Opportunity (FEO) requires that social positions, such as jobs, be formally open and meritocratically allocated, but, in addition, each individual is to have a fair chance to attain these positions. that considers these inequalities as acceptable : natural\_inequality, talent\_inequality, preference\_inequality and that considers these inequalities as unacceptable : ; 'socioeconomic\_inequality, treatment\_inequality')

FIGURE C.16: Component : Social Fairness Notions - Reviewing fairness philosophies (generated from mappings)

Specify which fairness definition is applicable to the model. The requirement Mandatory means that it is a must have, Experiment means that the definition can be explored and Not Applicable means that the definition does not need to be considered at all. If you set a fairness definition as mandatory, or experimentation, then the model developer will be told to ensure that these definitions are being met, or were explored.

### Equal Odds

- Description** : Equalized odds is satisfied provided that no matter whether an applicant is a Lilliputian or a Brobdingnagian, if they are qualified, they are equally as likely to get admitted to the program, and if they are not qualified, they are equally as likely to get rejected.,
- Intuition** : Suppose Group A and Group B contain two features, where one feature is sensitive (i.e. you want to make sure you are being fair towards this feature, e.g. Group A can be Male, and Group B can be female). Among applicants who are creditworthy and would have repaid their loans, both Group A and Group B applicants should have similar rate of their loans being approved

Select whether Equal Odds is mandatory, exploratory or not applicable

Mandatory  Not applicable

Please state your motivation

FIGURE C.17: Component : Fairness Definitions - Specifying fairness definitions

# Bibliography

- [1] *Aequitas - The Bias Report*. URL: <http://aequitas.dssg.io/>.
- [2] *Aequitas – Center for Data Science and Public Policy*. URL: <http://www.datasciencepublicpolicy.org/projects/aequitas/>.
- [3] *AI Fairness 360*. URL: <https://aif360.mybluemix.net/>.
- [4] Saleema Amershi et al. “Software Engineering for Machine Learning: A Case Study”. In: (). URL: <https://docs.microsoft.com/en-us/azure/devops/learn/devops-at-microsoft/>.
- [5] Yasuhito Arimoto, Masaki Nakamura, and Kokichi Futatsugi. “Toward a Domain Description with CafeOBJ”. In: ().
- [6] Solon Barocas et al. “Electronic Privacy Information Center”. In: *J.D* (2011). DOI: [10.15779/Z38BG31](https://doi.org/10.15779/Z38BG31). URL: <http://dx.doi.org/10.15779/Z38BG31>.
- [7] Rabi Narayan Behera et al. “A Survey on Machine Learning: Concept, Algorithms and Applications Machine Learning View project International Journal of Innovative Research in Computer and Communication Engineering A Survey on Machine Learning: Concept, Algorithms and Applications”. In: *Article in International Journal of Innovative Research in Computer* (2017). ISSN: 2320-9798. DOI: [10.15680/IJIRCCE.2017](https://doi.org/10.15680/IJIRCCE.2017). URL: [www.ijircce.com](http://www.ijircce.com).
- [8] Richard Berk et al. “A Convex Framework for Fair Regression”. In: (2017).
- [9] Richard Berk et al. “Fairness in Criminal Justice Risk Assessments: The State of the Art.” in: <https://doi.org/10.1177/0049124118782533> 50.1 (July 2018), pp. 3–44. DOI: [10.1177/0049124118782533](https://doi.org/10.1177/0049124118782533). URL: <https://journals.sagepub.com/doi/10.1177/0049124118782533>.
- [10] Reuben Binns. “Fairness in Machine Learning: Lessons from Political Philosophy”. In: *Proceedings of Machine Learning Research* 81 (2018), pp. 1–11.
- [11] Sarah Bird Facebook et al. “Fairness-Aware Machine Learning: Practical Challenges and Lessons Learned”. In: (2019). DOI: [10.1145/3289600.3291383](https://doi.org/10.1145/3289600.3291383). URL: <https://doi.org/10.1145/3289600.3291383>.
- [12] Alan W. Brown. “Model driven architecture: Principles and practice”. In: *Software and Systems Modeling* (Aug. 2004). DOI: [10.1007/S10270-004-0061-2](https://doi.org/10.1007/S10270-004-0061-2).
- [13] Yuriy Brun and Alexandra Meliou. “Software Fairness”. In: (2018). DOI: [10.1145/3236024.3264838](https://doi.org/10.1145/3236024.3264838). URL: <https://doi.org/10.1145/3236024.3264838>.
- [14] Joy Buolamwini and Timnit Gebru. *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*. Jan. 2018. URL: <https://proceedings.mlr.press/v81/buolamwini18a.html>.
- [15] Simon Caton and Christian Haas. “Fairness in Machine Learning: A Survey”. In: (Oct. 2020). URL: <http://arxiv.org/abs/2010.04053>.
- [16] Peter. Checkland. “Systems thinking, systems practice”. In: (1981), p. 330.

- [17] Betty H.C. Cheng and Joanne M. Atlee. "Research directions in requirements engineering". In: *FoSE 2007: Future of Software Engineering* (2007), pp. 285–303. DOI: [10.1109/FoSE.2007.17](https://doi.org/10.1109/FoSE.2007.17).
- [18] Edmund M. Clarke et al. "Formal methods: State of the art and future directions". In: *ACM Computing Surveys* 28.4 (1996), pp. 626–643. DOI: [10.1145/242223.242257](https://doi.org/10.1145/242223.242257).
- [19] Range Cleaveland et al. "Strategic directions in concurrency research". In: *ACM Computing Surveys* 28.4 (1996), pp. 607–625. DOI: [10.1145/242223.242252](https://doi.org/10.1145/242223.242252).
- [20] Jane Cleland-Huang et al. "Goal-Centric Traceability for Managing Non-Functional Requirements". In: *Proceedings of the 27th international conference on Software engineering - ICSE '05* (2004). DOI: [10.1145/1062455](https://doi.org/10.1145/1062455).
- [21] Ayala Cohen. "Comparison of correlated correlations". In: *Statistics in Medicine* 8.12 (Dec. 1989), pp. 1485–1495. ISSN: 1097-0258. DOI: [10.1002/SIM.4780081208](https://doi.org/10.1002/SIM.4780081208). URL: <https://onlinelibrary.wiley.com/doi/full/10.1002/sim.4780081208><https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.4780081208><https://onlinelibrary.wiley.com/doi/10.1002/sim.4780081208>.
- [22] Jason A. Colquitt. "On the dimensionality of organizational justice: A construct validation of a measure". In: *Journal of Applied Psychology* 86.3 (2001), pp. 386–400. DOI: [10.1037/0021-9010.86.3.386](https://doi.org/10.1037/0021-9010.86.3.386). URL: [/record/2001-06715-002](https://record/2001-06715-002).
- [23] Sam Corbett-Davies et al. *The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning* \*. Tech. rep. 2018.
- [24] Elizamary De Souza Nascimento et al. "Understanding Development Process of Machine Learning Systems: Challenges and Solutions". In: *International Symposium on Empirical Software Engineering and Measurement 2019-Septemer* (Sept. 2019). DOI: [10.1109/ESEM.2019.8870157](https://doi.org/10.1109/ESEM.2019.8870157).
- [25] Christian Denger, Daniel M Berry, and Erik Kamsties. "Higher Quality Requirements Specifications through Natural Language Patterns". In: (2003).
- [26] Cynthia Dwork et al. "Fairness through awareness". In: *ITCS 2012 - Innovations in Theoretical Computer Science Conference* (2012), pp. 214–226. DOI: [10.1145/2090236.2090255](https://doi.org/10.1145/2090236.2090255).
- [27] Jonathan Edwards, Daniel Jackson, and Emina Torlak. "A type system for object models". In: (2004), p. 189. DOI: [10.1145/1029894.1029921](https://doi.org/10.1145/1029894.1029921).
- [28] *Exploring or Exploiting? Social and Ethical Implications of Autonomous Experimentation in AI* by Sarah Bird, Solon Barocas, Kate Crawford, Fernando Diaz, Hanna Wallach :: SSRN. URL: [https://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=2846909](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2846909).
- [29] *Fairness in Machine Learning: A Survey* | DeepAI. URL: <https://deepai.org/publication/fairness-in-machine-learning-a-survey>.
- [30] Ali Farahani et al. "On Adaptive Fairness in Software Systems". In: ().
- [31] *Firebase*. URL: [https://firebase.google.com/?gclid=CjwKCAjwgvviIBhBkEiwA10D2j55\\_AUGJmV5c10TB4yCQNVsXniMSQE\\_Qoe6G7EgXkvATkHy7j66eahoCh5kQAvD\\_BwE&gclidsrc=aw.ds](https://firebase.google.com/?gclid=CjwKCAjwgvviIBhBkEiwA10D2j55_AUGJmV5c10TB4yCQNVsXniMSQE_Qoe6G7EgXkvATkHy7j66eahoCh5kQAvD_BwE&gclidsrc=aw.ds).
- [32] Ariel Fuxman et al. "Specifying and analyzing early requirements in Tropos". In: *Requirements Engineering* 9.2 (May 2004), pp. 132–150. DOI: [10.1007/S00766-004-0191-7](https://doi.org/10.1007/S00766-004-0191-7).



- [33] Sainyam Galhotra, Yuriy Brun, and Alexandra Meliou. "Fairness Testing: Testing Software for Discrimination". In: (2017). DOI: [10.1145/3106237.3106277](https://doi.org/10.1145/3106237.3106277). URL: <https://doi.org/10.1145/3106237.3106277>.
- [34] Jean Garcia-Gathright Spotify Somerville, Aaron Springer, and Spotify San Francisco. "Assessing and Addressing Algorithmic Bias-But Before We Get There Henrieee Cramer". In: (2018).
- [35] Jerald Greenberg. "A Taxonomy of Organizational Justice Theories". In: 12.1 (1987), pp. 9–22.
- [36] Carl A. Gunter et al. "Reference model for requirements and specifications". In: *IEEE Software* 17.3 (May 2000), pp. 37–43. DOI: [10.1109/52.896248](https://doi.org/10.1109/52.896248).
- [37] J. G. Hall and L. Rapanotti. "A reference model for requirements engineering". In: *Proceedings of the IEEE International Conference on Requirements Engineering* 2003-January (2003), pp. 181–187. DOI: [10.1109/ICRE.2003.1232749](https://doi.org/10.1109/ICRE.2003.1232749).
- [38] Kenneth Holstein et al. "Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need?" In: *ACM Reference Format: Kenneth Holstein ()*, p. 16. DOI: [10.1145/3290605.3300830](https://doi.org/10.1145/3290605.3300830). URL: <https://doi.org/10.1145/3290605.3300830>.
- [39] Kenneth Holstein et al. "Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need". In: *ACM Reference Format: Kenneth Holstein ()*, p. 16. DOI: [10.1145/3290605.3300830](https://doi.org/10.1145/3290605.3300830). URL: <https://doi.org/10.1145/3290605.3300830>.
- [40] George Casper Homans. "The Humanities and the Social Sciences:" in: <http://dx.doi.org/10.1177/000276426100400802> 4.8 (Nov. 2016), pp. 3–6. DOI: [10.1177/000276426100400802](https://doi.org/10.1177/000276426100400802). URL: <https://journals.sagepub.com/doi/abs/10.1177/000276426100400802>.
- [41] *How We Analyzed the COMPAS Recidivism Algorithm — ProPublica*. URL: <https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm>.
- [42] Steven de Jong and Karl Tuyls. "Human-inspired computational fairness". In: *Autonomous Agents and Multi-Agent Systems 2010* 22:1 22.1 (Feb. 2010), pp. 103–126. ISSN: 1573-7454. DOI: [10.1007/S10458-010-9122-9](https://doi.org/10.1007/S10458-010-9122-9). URL: <https://link.springer.com/article/10.1007/s10458-010-9122-9>.
- [43] Matthew Kay, Cynthia Matuszek, and Sean A. Munson. "Unequal representation and gender stereotypes in image search results for occupations". In: *Conference on Human Factors in Computing Systems - Proceedings* 2015-April (Apr. 2015), pp. 3819–3828. DOI: [10.1145/2702123.2702520](https://doi.org/10.1145/2702123.2702520).
- [44] Ron Kohavi. "Guest editors' introduction: On applied research in machine learning". In: ().
- [45] Claudio Lazo. *Towards Engineering AI Software for Fairness: A framework to help design fair, accountable and transparent algorithmic decision-making systems*. 2020. URL: <https://repository.tudelft.nl/islandora/object/uuid%3Ac7dc661e-c3f9-4986-bc54-c903aaddbc68>.
- [46] Michelle Seng Ah Lee, Luciano Floridi, and Jatinder Singh. "Formalising trade-offs beyond algorithmic fairness: lessons from ethical philosophy and welfare economics". In: *SSRN Electronic Journal* (July 2020). DOI: [10.2139/SSRN.3679975](https://doi.org/10.2139/SSRN.3679975). URL: <https://papers.ssrn.com/abstract=3679975>.

- [47] Emmanuel Letier and Axel van Lamsweerde. "Agent-based tactics for goal-oriented requirements elaboration". In: (2002), p. 83. DOI: [10.1145/581339.581353](https://doi.org/10.1145/581339.581353).
- [48] Gerald S. Leventhal and James W. Michaels. "Extending the equity model: Perception of inputs and allocation of regard as a function of duration and quantity of performance". In: *Journal of Personality and Social Psychology* 12.4 (Aug. 1969), pp. 303–309. DOI: [10.1037/H0027798](https://doi.org/10.1037/H0027798). URL: [/record/1969-15729-001](https://record/1969-15729-001).
- [49] Giuliano Lorenzoni et al. "Machine Learning Model Development from a Software Engineering Perspective: A Systematic Literature Review". In: (Feb. 2021). URL: <https://arxiv.org/abs/2102.07574v1>.
- [50] Ninareh Mehrabi et al. "A Survey on Bias and Fairness in Machine Learning". In: (Aug. 2019). URL: <http://arxiv.org/abs/1908.09635>.
- [51] Bashar Nuseibeh, Jeff Kramer, and Anthony Finkelstein. "ViewPoints: meaningful relationships are difficult!" In: (). URL: <http://www.xlinkit.com..>
- [52] (PDF) *Classification of Research Efforts in Requirements Engineering*. URL: [https://www.researchgate.net/publication/220565934\\_Classification\\_of\\_Research\\_Efforts\\_in\\_Requirements\\_Engineering](https://www.researchgate.net/publication/220565934_Classification_of_Research_Efforts_in_Requirements_Engineering).
- [53] Dana Pessach. "Algorithmic Fairness". In: ().
- [54] Yusuf Pisan. "Extending Requirement Specifications Using Analogy". In: (). URL: <http://www.ics.mq.edu.au/~ypisan/>.
- [55] Emily Pronin. "Perception and misperception of bias in human judgment". In: *Trends in Cognitive Sciences* 11.1 (Jan. 2007), pp. 37–43. DOI: [10.1016/J.TICS.2006.11.001](https://doi.org/10.1016/J.TICS.2006.11.001). URL: [/record/2006-23397-007](https://record/2006-23397-007).
- [56] Junfei Qiu et al. "A survey of machine learning for big data processing". In: (2016). DOI: [10.1186/s13634-016-0355-x](https://doi.org/10.1186/s13634-016-0355-x).
- [57] John Rawls. "A theory of justice". In: (1999), p. 538. URL: [https://books.google.com/books/about/A\\_Theory\\_of\\_Justice.html?id=kvpby7HtAe0C](https://books.google.com/books/about/A_Theory_of_Justice.html?id=kvpby7HtAe0C).
- [58] *Requirements trawling: techniques for discovering requirements* | Elsevier Enhanced Reader. URL: <https://reader.elsevier.com/reader/sd/pii/S1071581901904811?token=656A20C462C725C69491E1CF7BE92A95C093A8116E066EC14630DBBFFDD3465725C53FF55E2&originRegion=eu-west-1&originCreation=20210722090036>.
- [59] *Research on Scientific Reasoning - Anderson - 1999 - Journal of Research in Science Teaching - Wiley Online Library*. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/%28SICI%291098-2736%28199909%2936%3A7%3C751%3A%3AAID-TEA1%3E3.0.CO%3B2-R>.
- [60] Suzanne. Robertson and James Robertson. "Mastering the requirements process : getting requirements right". In: (2013).
- [61] Boris Ruf, Chaouki Boutharouite, and Marcin Detyniecki. "Getting Fairness Right: Towards a Toolbox for Practitioners Author Keywords CCS Concepts". In: (2020). DOI: [10.1145/3334480.XXXXXX](https://doi.org/10.1145/3334480.XXXXXX). URL: <https://doi.org/10.1145/3334480.XXXXXX>.
- [62] Kailash Karthik Saravanakumar. "The Impossibility Theorem of Machine Fairness – A Causal Perspective". In: (July 2020). URL: <https://arxiv.org/abs/2007.06024v2>.
- [63] Andrew D Selbst et al. "Fairness and Abstraction in Sociotechnical Systems". In: (2019). DOI: [10.1145/3287560.3287598](https://doi.org/10.1145/3287560.3287598). URL: <https://doi.org/10.1145/3287560.3287598>.

- [64] Helen Sharp, Anthony Finkelstein, and Galal Galal. "Stakeholder Identification in the Requirements Engineering Process". In: ().
- [65] Andres Silva. "Requirements, Domain and Specifications: A Viewpoint-based Approach to Requirements Engineering". In: (2002).
- [66] Froukje Sleeswijk Visser. "Bringing the everyday life of people into design". In: (). URL: [www.contextmapping.com](http://www.contextmapping.com).
- [67] Streamlit. URL: <https://streamlit.io/>.
- [68] Alistair Sutcliffe, Stephen Fickas, and McKay Moore Sohlberg. "PC-RE: A method for personal and contextual requirements engineering with some experience". In: *Requirements Engineering* 11.3 (June 2006), pp. 157–173. DOI: [10.1007/S00766-006-0030-0](https://doi.org/10.1007/S00766-006-0030-0).
- [69] *Tell Me a Story - Northwestern University Press*. URL: <https://nupress.northwestern.edu/9780810113138/tell-me-a-story/>.
- [70] Jens Damgaard Thaysen and Andreas Albertsen. "When bad things happen to good people: Luck egalitarianism and costly rescues". In: *Politics, Philosophy and Economics* 16.1 (Feb. 2017), pp. 93–112. DOI: [10.1177/1470594X16666017](https://doi.org/10.1177/1470594X16666017). URL: <https://philpapers.org/rec/THAWBT-2>.
- [71] "The Triptych Paradigm". In: *Software Engineering* 3 (June 2006), pp. 3–51. DOI: [10.1007/3-540-33653-2\\_{\\\_}1](https://doi.org/10.1007/3-540-33653-2_{\_}1).
- [72] Sebastian Uchitel, Jeff Kramer, and Jeff Magee. "Behaviour Model Elaboration using Partial Labelled Transition Systems". In: (2000).
- [73] Sebastian Uchitel, Jeff Kramer, and Jeff Magee. "Detecting implied scenarios in message sequence chart specifications". In: (2001), p. 74. DOI: [10.1145/503209.503220](https://doi.org/10.1145/503209.503220).
- [74] Sebastian Uchitel, Jeff Kramer, and Jeff Magee. "Incremental elaboration of scenario-based specifications and behavior models using implied scenarios". In: *ACM Transactions on Software Engineering and Methodology* 13.1 (Jan. 2004), pp. 37–85. DOI: [10.1145/1005561.1005563](https://doi.org/10.1145/1005561.1005563).
- [75] Sebastian Uchitel, Jeff Kramer, and Jeff Magee. "Negative scenarios for implied scenario elicitation". In: (2002), p. 109. DOI: [10.1145/587051.587069](https://doi.org/10.1145/587051.587069).
- [76] Axel Van Lamsweerde. "Handling obstacles in goal-oriented requirements engineering". In: *IEEE Transactions on Software Engineering* 26.10 (Oct. 2000), pp. 978–1005. DOI: [10.1109/32.879820](https://doi.org/10.1109/32.879820).
- [77] Sahil Verma and Julia Rubin. "Fairness Definitions Explained". In: *IEEE/ACM International Workshop on Software Fairness* 18 (2018). DOI: [10.1145/3194770.3194776](https://doi.org/10.1145/3194770.3194776). URL: <https://doi.org/10.1145/3194770.3194776>.
- [78] N I Versity, Ian Sommerville, and Pete Sawyer. "LANCASTER U Computing Department Viewpoints: Principles, Problems and a Practical Approach to Requirements Engineering RUNNING TITLE: VIEWPOINTS FOR REQUIREMENTS ENGINEERING Viewpoints: principles, problems and a practical approach to requirements engineering". In: (). URL: [http://www.comp.lancs.ac.uk/computing/research/cseg/http://www.comp.lancs.ac.uk/computing/research/cseg/97\\_rep.html](http://www.comp.lancs.ac.uk/computing/research/cseg/http://www.comp.lancs.ac.uk/computing/research/cseg/97_rep.html).
- [79] Mihaela Vorvoreanu et al. *From Gender Biases to Gender-Inclusive Design: An Empirical Investigation*. May 2019. DOI: [10.1145/3290605.3300283](https://doi.org/10.1145/3290605.3300283). URL: <https://www.microsoft.com/en-us/research/publication/from-gender-biases-to-gender-inclusive-design-an-empirical-investigation/>.

- [80] *What-If Tool*. URL: <https://pair-code.github.io/what-if-tool/>.