



MASTER THESIS

<AI & EQUALITY> A HUMAN RIGHTS TOOLBOX

Author:
Sofia KYPRAIOU

Supervisors:
Prof. Daniel GATICA-PEREZ
Caitlin KRAFT-BUCHMAN

IDIAP RESEARCH INSTITUTE (IDIAP)
SCHOOL OF COMPUTER AND COMMUNICATION SCIENCES (IC)
WOMEN AT THE TABLE

Lausanne, August 2021

Life is pain
au chocolate.
— unknown

To my parents...

Acknowledgements

First, I would like to thank Caitlin Kraft-Buchman, CEO and Founder of Women at the Table, for all the guidance and help she gave me. I'm truly grateful for her trust in me in this project. I'm thankful for all the discussions we had, the opportunities, the work we did. It helped me improve intellectually and raise my social awareness.

I'd also like to thank my thesis supervisor, EPFL Professor Daniel Gatica-Perez, who leads the Social Computing Group at Idiap. The interventions and insights were always on point, and I'm thankful for all the constructive suggestions and recommendations you gave.

I want to thank our human rights expert, human rights officer at the Women's Human Rights and Gender Section, the Office of the United Nations High Commissioner for Human Rights (UN Human Rights – OHCHR), Asako Hattori. I have learned so many important concepts from our weekly meetings when designing the toolbox. You are a crucial part of this approach, and you've contributed enormously with ideas and concepts from human rights. I'm deeply thankful for that.

I'd also like to thank Professor Susan Leavy, Assistant Professor in the School of Information and Communication at the University College Dublin. I'm very grateful for agreeing to host the workshop at UCD and for all the help with the organisation and selection of participants. Thank you.

Finally, I'd like to thank all the participants in the two workshops who were part of this methodology and completed a survey. And a special mention to the participant who agreed to stay for the long interview.

Lausanne, 27 August 2021

S. K.

Abstract

There is a gap in computer science education where scientists have solid technical backgrounds but often lack substantial human rights knowledge. As the data science field evolves quickly, many universities do not have time to adjust their curriculum and create interdisciplinary courses. In addition, many approaches focus on tech ethics; however, ethics and fairness lack a unique definition - they are open to different interpretations.

Universities have a critical role in meeting these ethical concerns. By integrating human values into computing technology, universities can educate scientists and engineers to work for the public good, creating technology in accordance with human values.

This study investigates how to integrate a human rights-based approach based on the International Human Rights framework with current concepts of fairness for designing an educational tool for computer scientists. It converts voluntary promises of ethical behaviour into compulsory requirements for compliance with established legislation.

The methodology *<AI & Equality> A Human Rights Toolbox* was developed in conjunction with experts for human rights through the Office of the United Nations High Commissioner for Human Rights (OHCHR) and includes a workshop consisting of a Human Rights module and code, outreach and community plan incorporating human rights concepts with data science. To evaluate and validate the methodology, two workshops were conducted with 12 participants. Our analysis showed that the students responded well, with a marked increase in their awareness of human rights principles. They improved their ability to identify and analyze how gender, racial and other bias occurred or can occur in the research, design, and development of AI, in addition to their ability to identify and use tools and techniques to mitigate bias in AI.

Key words: human rights, education, fairness, human rights-based approach, machine learning

Contents

Acknowledgements	i
Abstract	iii
List of figures	vii
List of tables	ix
1 Introduction	1
1.1 Problem definition and motivation	1
1.2 Aim and research objectives	2
1.3 Research approach	3
1.4 Thesis structure	4
2 Literature Review	7
2.1 Research on fairness	7
2.1.1 Limitations	9
2.2 AI Fairness Tools: Principles and Practice	9
2.2.1 Ethics Principles	9
2.2.2 Documentation Tools	10
2.2.3 Software Toolkits	10
2.2.4 Limitations	13
2.3 Teaching of tech ethics in university curriculum	14
2.3.1 Limitations	15
2.4 Human Rights frameworks	16
2.4.1 Limitations	17
3 Methodology	19
3.1 Overall approach	20
3.1.1 Goal, learning objectives and outcomes	20
3.1.2 Beneficiaries	21
3.1.3 Workshop structure	21
3.2 Human Rights Module	22
3.2.1 Objectives	22
3.2.2 Structure	22

3.2.3	Content	23
3.3	Applied Research	26
3.3.1	Objectives	26
3.3.2	Structure	27
3.4	Practicum	27
3.4.1	Objectives	27
3.4.2	Structure	28
3.4.3	Roadmap	28
3.4.4	Introduction to fairness	29
3.4.5	Baseline Model	36
3.4.6	Pre-processing	41
3.4.7	In-processing	44
3.4.8	Post-processing	45
4	Evaluation of the framework	49
4.1	Testing the methodology: Workshop #2 - EPFL	50
4.2	Participants	50
4.3	Questionnaire	51
4.4	Feedback	52
4.5	Specifications	53
5	Validation of the framework	55
5.1	Validating the methodology: Workshop #3 - UCD	55
5.2	Participants	55
5.3	Results of Questionnaire	56
5.4	Analysis of results	57
6	Discussion	61
6.1	Summary	61
6.2	Contribution	63
6.3	Limitations of the methodology	63
6.4	Recommendations for future work	64
7	Conclusion	67
A	An appendix	69
A.1	Biases	69
A.2	Questionnaires	73
	Bibliography	84

List of Figures

3.1	Equality under international human rights law	25
3.2	A Human Rights-based approach	26
3.3	Example slides from workshop	31
3.4	Demographic parity with laziness	32
3.5	Equalized odds	33
3.6	Exploratory demographics plots	39
3.7	Exploratory target plots	39
3.8	Evaluation of the baseline model	40
3.9	Bias in the data	42
3.10	Evaluation of the pre-processing model	43
3.11	Evaluation of the in-processing model	45
3.12	Evaluation of the post-processing model	47
4.1	Poster for the workshop at EPFL	50
4.2	Demographic plots and student status of the EPFL workshop participants . . .	51
5.1	Demographic plots and student status of the UCD workshop participants . . .	56

List of Tables

2.1	Software toolkits for data scientists and practitioners addressing AI Fairness. . .	12
3.1	Demographic Parity vs Equalized Odds: Job application example	35
4.1	Places where we presented the workshop <AI & Equality> A Human Rights Toolbox and registered participants	50
5.1	The evaluation of the objectives by the participants from Workshop #3 at UCD	56
A.1	Example of different types of bias (Funding and Historic)	70
A.2	Example of different types of bias (Sampling and Temporal)	71
A.3	Example of different types of bias (Representation)	72
A.4	The survey used for the evaluation of the workshop	73

Acronyms

EPFL École Polytechnique Fédérale de Lausanne. 3, 4, 15, 19, 22, 27, 49, 50, 55, 62, 63

HRBA Human Rights-Based Approach. 4, 17–20, 22, 24, 25, 28

NGO Non-Governmental Organisation. 3, 19, 22

OHCHR Office of the United Nations High Commissioner for Human Rights. iii, 3, 17, 19, 22, 23, 51, 52, 63

UCD University College Dublin. 4, 49, 50, 55, 62, 65, 66

UDHR Universal Declaration of Human Rights. 16, 19, 23

UN United Nations. 16, 17, 19, 24

W@tt Women at the Table. 3, 19, 22, 66

1 Introduction

1.1 Problem definition and motivation

The rise of new technologies and the growth of machine learning algorithms and applications have transformed most people's lives. They play a significant part in realizing positive social and economic advances.

When these technologies are designed and engineered to serve humans, they contribute to economic growth and employment creation (Szczepański, 2019). They improve equality and participation, empower civil and human rights activism (Rowley, 2018), and open up new opportunities for innovation in the public and social sector (Misuraca & Viscusi, 2020). They aid to hold institutions, companies, and governments accountable for their actions (Schaake, 2020) and improve the efficiency of science.

Nevertheless, at the same time, the misuse, the wrong design, the limitations, and the bias of these very same innovations have become the focus of criticism for many researchers. Reports reveal the algorithmic bias in every aspect of modern life. It has a far-reaching direct impact on work environments (Ajunwa & Greene, 2019) (Mayer, 2018), with the implementation of AI algorithms being extensively used in recruitment, assessing candidates' performance at interview (Harwell, 2019) or sometimes showing, for example, bias against women (Dastin, 2018).

Machine learning applications are also used in college recruitment (Jaquette & Salazar, 2018), college admissions (Waters & Miikkulainen, 2014), or predictors for student's success (Feathers, 2021). All of the above approaches have currently been abandoned as they exacerbate existing inequalities in the field.

All of these applications affect our lives directly. If they are not designed by incorporating concepts of human rights and fairness, they can negatively affect our society.

The discussion of various perspectives and viewpoints of fairness has already been happening from the 1970s (Poon, 2012) (Ochigame, 2020). In many modern fairness concepts, people have

identified the potential harms that come from big uncurated data. Author and mathematician Cathy O’Neil wrote about the societal impact of algorithms, exploring how big data increases inequality and threatens democracy (O’Neil, 2016). Other authors research and explore how algorithms and bias are far from neutral, as they replicate and reinforce biased beliefs into search engines results (Noble, 2018) and how technology influences civil, political, and human rights and affects economic equity (Eubanks, 2018). Another way to explore the algorithmic bias is through the lens of power, analyzing algorithms through a feminist theory (D’Ignazio & Klein, 2020) or politics, exploitation of natural resources and labour (Crawford, 2021).

There is a growing realization that a critical path will be to address how and what data/computer science/machine learning students are taught about human rights. This path may either alleviate (or exacerbate) algorithmic discrimination and bias in data, models, and neural networks to aid in designing inclusive human rights-based systems.

The new generation of researchers tasked with creating new algorithmic systems have solid technical backgrounds but lack substantial human rights knowledge or frameworks to use this technical knowledge as “AI for Social Good”.

As the data science/machine learning field is evolving quickly, many universities do not have time to adjust the material and adapt to the changes, if they even think that human rights are an appropriate subject to be taught in this domain. There is a deficit of bandwidth, expertise, and knowledge regarding the relationship between “technique” and human rights, which may leave another generation of scientists disempowered to leverage their education to make the impact the world needs.

1.2 Aim and research objectives

The aim of this study is to investigate:

RQ: How can International Human Rights frameworks be integrated with current concepts of fairness for the design of an educational tool for computer scientists?

The following objectives help to address this research question :

Research Objective 1: Study current approaches on ethics and fairness in machine learning and examine their main limitations

Research Objective 2: Propose, design, and build an educational tool for computer scientists

- Propose how to combine different practical fairness tools in a single Jupyter notebook
- Examine how to incorporate a human rights-based approach in the fairness pipeline

- a Pre-processing (data)
- b In-processing (model)
- c Post-processing (evaluation)

1.3 Research approach

Our solution to help bridge the gap in the data science and human rights education is the workshop, outreach and community plan embedded within *<AI & Equality> A Human Rights Toolbox*¹. The Toolbox is based on an idea of the NGO Women at the Table (W@tt), and its collaboration with the Office of the United Nations High Commissioner for Human Rights (OHCHR) and its Women's Human Rights and Gender Section. Piloted at EPFL through workshops, *<AI & Equality> A Human Rights Toolbox* addresses the problem head-on in language computer science students can both understand and use. It proves a basic human rights workshop blended with tools to directly and immediately see how human rights principles can be applied and thought about analytically in code. This Toolbox is the first and to date only foray of OHCHR into the world of university computer science students in order to jumpstart a conversation about a human rights-based approach being the baseline from which we should create new algorithms and new models.

We focus on human rights instead of ethics, which are often better defined and measurable, as most are defined under international or national law. According to the Office of the United Nations High Commissioner for Human Rights (OHCHR) "Human rights are rights we have because we exist as human beings. These universal rights are inherent to us all, regardless of nationality, sex, national or ethnic origin or any other status" (The United Nations, 1948). They are based on international law and provide an ethical lens that exceeds national and cultural borders. Human rights put people in the centre of decision-making and can assess and address any unintentional harm.

We created a light touch Human Rights Module pitched to computer science students in collaboration with OHCHR and added a discussion of current applied research in the domain of fairness. Then we explored the concepts with an in-depth Jupyter notebook that takes the Human Rights module and marries it in discussion to concrete examples and exercises that reveal how human rights principles are at interplay with the code. The Jupyter notebook investigates how code can be de-biased and improved to support Human Rights principles of non-discrimination and equality, participation, and accountability (UNSDG Human Rights Working Group, 2003).

We tested our methodology through an iterative series of workshops, and we found that the students responded well, with a marked increase in their awareness of human rights principles. They improved their ability to identify and analyze how gender, racial and other bias occurred or can occur in the research, design, and development of AI, in addition to their ability to

¹Website for the project: <https://aiequalitytoolbox.com/>

identify and use tools and techniques to mitigate bias in AI.

1.4 Thesis structure

After the introduction discussed above, the thesis is divided into the consequent chapters, described in detail in the following sections.

Chapter 2 - Literature Review

Chapter 2 reviews the current research on fairness in machine learning and the limitations of using ethics in a technical concept. We continue by presenting the available checklists, principles, documentation tools, and software toolkits recently deployed to guide a more ethical approach to AI. We highlight the limitations that prevent these tools from being widely used. Then, we explore how different universities have attempted to include tech ethics in their curriculum and the difficulties of these approaches. Finally, we introduce the Human Rights-Based Approach (HRBA) to deal with the uncertainty of ethics.

Chapter 3 - Methodology

Chapter 3 outlines the methodological approach of the workshop that we created. It explains and supports the research methods we used, and it showcases the main pillars of the workshop and the Jupyter notebook: The human rights module and the practicum.

Chapter 4 - Evaluation

The chapter describes the procedures undertaken during the evaluation of methodology at the workshop presented at EPFL on March 26th, 2021. Then, we present the key findings that illustrate the workshop's success, but also the difficulty participants had in digesting a large amount of new information.

Chapter 5 - Validation

Chapter 5 presents the procedures and the improvements we made based on the specifications of the previous evaluation. We then present the reception from a workshop at University College Dublin (UCD) on May 20th, 2021. Our goal was to obtain additional feedback regarding the content and issues related to the usability of the delivered information.

Chapter 6 - Discussion

Chapter 6 summarises the thesis and reviews the main contributions of this research. We present the limitations of our work, and then we conclude with opportunities for future

research and our proposals for expanding the work of *<AI & Equality>*.

Chapter 7 - Conclusion

Finally, the last chapter summarizes and concludes the research.

2 Literature Review

This thesis explores how we can integrate a human rights-based approach, instead of ethics, to the fairness pipeline most commonly used in machine learning literature and combine this knowledge in an educational Jupyter notebook.

First, we explore how the concept of ethics and different definitions of fairness are translated as a mathematical property of machine learning algorithms and discuss different definitions of what makes an algorithm fair. Next, we discuss the limitations of using ethics given their socio-cultural context and how their translation into a technical solution removes the complexity of the real-world problems that algorithms are trying to solve.

The need to tackle ethics issues in technologies has led to the creation and development of heuristic tools, checklists, and software toolkits. However, although this is an advance on the right track, heuristic tools are not adequately integrated into the design and evaluation of an algorithm.

Finally, we see why a human rights-based approach from settled international human rights law provides a better framework for dealing with complex, global issues of algorithmic accountability and fairness.

2.1 Research on fairness

Today, machine learning algorithms are used to support decision-making. The algorithms often rely on multiple datasets to learn a model, then applied to other use cases. Research continues to reveal the risks of this form of algorithmic decision-making and the risks it presents augmenting human biases, especially for those already marginalized. The general assumption is that the more data is used, the more precise the algorithm and its predictions are. When using a large amount of data, it contains many correlations between the different features and variables of the dataset. However, not all correlations imply causality, and more data does not necessarily mean better. Because no matter how large the dataset is, it still only remains a snapshot of reality.

In order to deal with, understand, and correct algorithmic biases, researchers are trying to identify, measure and improve algorithmic fairness when using AI algorithms. But what do we mean when we want a fair and non-discriminatory machine learning model? How can this be achieved?

Defining what is fair, however, is far from a straightforward task. Fairness can mean different things to different people, and questions like “is fair if everyone is treated equally, does it minimize harms?” are derived from ethics and political philosophy (Binns, 2021). Fairness is an ethical concept which refers to plural conceptions of justice between individuals. This concept is at the heart of social science research (GOFF, 1983) (Birnbacher, 1999) (Kitchener & Kitchener, 2009), and the difficulty to find a general definition is obvious: fairness is based on an ethical value judgment, and its application will vary according to cultures, religions, political systems.

To use fairness as a concept in machine learning models, it needs to be translated into a mathematical notion. However, formalizing a notion of fairness in machine learning and putting it in mathematical terms is open to different interpretations. Different definitions have been proposed, using different assumptions based on definitions and concepts such as discrimination, justice, and equality.

Researchers have been creating mathematical modes and toolkits for fairness, explainability, transparency, and privacy. However, these methods tend to translate complex concepts into simplified and quantitative definitions. There are over twenty different fairness definitions, some of which are mutually incompatible (Friedler et al., 2016).

Authors (Barocas et al., 2019), (S. Mitchell et al., 2018) (Verma & Rubin, 2018) and survey papers (Caton & Haas, 2020), (Mehrabi et al., 2019) have collected the different definitions of fairness, comparing them and explaining their limitations and usages. What all authors agree is that there is no explicit consensus on which definition best fits individual situations.

The approaches to fairness vary from mathematical definitions of fairness (Hardt et al., 2016), (Chouldechova, 2016), (Zafar, Valera, Gomez Rodriguez, et al., 2017) to algorithmic approaches on how to “fix” datasets (Celis et al., 2021) and algorithms (Berk, Heidari, Jabbari, Joseph, et al., 2017).

Most mathematical definitions of fairness focus on mitigating performance between racial or gender groups (Agarwal et al., 2018), (Berk, Heidari, Jabbari, Kearns, et al., 2017), (Chouldechova et al., 2018), (Dwork et al., 2011), (Narayanan, 2018). Although they might work in a mathematical setup and improve the fairness metrics, these metrics face real challenges when deployed to the real world. Moreover, both public- (Veale et al., 2018) and private-sector (Holstein et al., 2019) practitioners found the quantitative definitions of fairness proposed by researchers incompatible with their work.

However, there is a shift in the research community, with the increased popularity and push to

make machine learning topics like fairness, ethics, and justice as “core” as other topics, like optimization. The need to reflect and explore is also shown in the recent work. The authors (Kasy & Abebe, 2021) drift away from the group and individual definitions of fairness, and instead of “fixing” the bias, they ask more profound questions about power and justice.

2.1.1 Limitations

Ethics is a socio-cultural concept, not just a technical one, and needs to be treated holistically. Going after solely technical solutions is an oversimplification and leads to the “techno solutionism” of computer science (Metcalf et al., 2019). “Techno-solutionism”, a term that became popular with the work of writer and researcher Evgeny Morozov, (Morozov, 2013), describes our need to jump into technological solutions as a quick and easy fix of complex real-world problems. As tempting as it seems to solve complicated issues with data, technology is made and programmed by humans, subject to the same prejudice and bias. Most complex real-world problems require complex real-world solutions.

These narrowly focused tech solutions can also lead to what is defined as “ethics washing”: “a rhetorical commitment to addressing AI ethics issues that are unsupported by concrete actions” (Bietti, 2019). It also diverts the attention from whether concrete measures are taken towards creating a world where AI serves “just as good for women, people of colour, or young people as it does for the white men who make up the majority of people making AI systems” (Johnson, 2019).

2.2 AI Fairness Tools: Principles and Practice

In recent years, many public and private organizations have published checklists, principles, and value statements to guide a more ethical design and implementation of AI systems to create a more unified process. The approaches can be separated into lists and frameworks that focus on fairness, transparency, security.

This section presents ethics principles, documentation, and software toolkits created to apply these principles in practice.

2.2.1 Ethics Principles

The UK Government released a Data Ethics Framework as guidance for public sector organizations (Digital & Office, 2018). The European Union’s High-Level Expert Group published ethical guidelines for responsible AI for organizations (E. U. H.-l. E. Group, 2019). ACM updated their Code of Ethics to be in line with the current discussions of ethics for professional practitioners (Anderson, 1992). In April 2021, the European Commission proposed the first-ever legal framework on AI, establishing standards on artificial intelligence (Commission, 2021). However, this framework is still in its early stages en route to being legislated.

Other checklists [(Cramer et al., 2019), (for Government Excellence, 2019), (Garage, 2019), (Vallor, 2019), (Loukides et al., 2018), (DrivenData, 2019)] have been designed to incorporate ethics into the development of a machine learning model. Different checklists relate to different stakeholders and are structured differently, some by principle and others by the stage of the AI development and its deployment lifecycle (Madaio et al., 2020). One of them, “DrivenData’s “Deon” (DrivenData, 2019) ethics checklist, also expands to a python package.

Currently, principles are too abstract for data science and machine learning practitioners. It is not easy to translate values into code. Despite their good intentions, if concrete actions and other mechanisms to make sure that practitioners make ethical decisions do not complement the AI ethics principles, then these principles will fail to achieve their goal. Just having a set of principles does not mean that these principles will be considered when making decisions daily (Loukides et al., 2018).

2.2.2 Documentation Tools

Detailed documentation of AI systems is critical for accountability and the successful implementation of AI principles. Documentation procedures provide structured information and make algorithms and their development more auditable at the dataset or model levels. They identify and anticipate risks before deployment along with several phases of the AI Lifecycle. Although there are currently no generally accepted standardized AI documentation procedures, several (Benjamin et al., 2019) have recently been developed.

Other lists come in the form of questions and focus on different parts of the machine learning pipeline. *Datasheets for Datasets* (Gebru et al., 2018), asks the creator/owner of a dataset to clarify diverse areas of issues: “Motivation, composition, collection process, preprocessing/-cleaning/labeling, uses, distribution, maintenance”. The project *The Dataset Nutrition Label* (Holland et al., 2018), using food nutrition labels as inspiration, is a framework to standardize datasets and data analysis before using a machine learning model. Similar documentation approaches on the datasets have been proposed for NLP datasets (Bender & Friedman, 2018). *Model cards for Model Reporting* (M. Mitchell et al., 2018) focuses on the machine learning model. *Factsheets* (Arnold et al., 2019) uses a more general approach and focuses on transparency and documentation of AI as a service.

2.2.3 Software Toolkits

Several fairness toolkits have recently emerged to make fairness methods and bias mitigation techniques widely accessible to more practitioners, using either open source or proprietary software solutions. Table 2.1 has a list of the most common fairness toolkits for data scientists. In this analysis, only open-sourced toolkits were evaluated.

Tool	Developer	Features
Aequitas ¹ (Saleiro et al., 2019)	University of Chicago Center for Data Science and Public Policy	bias audit toolkit for machine learning developers, used to audit the predictions of machine learning-based risk assessment tools.
AI Fairness 360 ² (Bellamy et al., 2018b)	IBM	toolkit to support examination, report, and mitigation of bias in machine learning models
AI Explainability 360 ³ (Arya et al., 2020)	IBM	toolkit with algorithms to understand how machine learning models predict labels
audit-ai ⁴ (Wilson et al., 2021)	pymetrics	library that implements fairness-aware machine learning algorithms and statistical tests to assess fairness
Deon ⁵	Driven Data	command-line tool that enables to add an ethics checklist to data science projects
Model Guardian	Deloitte	proprietary software using quantitative and qualitative methods to assure fairness in models
Fairlearn ⁶ (Bird et al., 2020)	Microsoft	open-source python toolkit with fairness metrics and algorithms
Fairness Flow	Facebook	technical toolkit to analyze how some types of AI models and labels perform across different groups
LinkedIn Fairness Toolkit (LiFT) ⁷ (Vasudevan & Kenthapadi, 2020)	LinkedIn	Scala/Spark library that measures fairness and mitigates bias in large-scale machine learning workflows
REVISE: REvealing Visual biasSEs (Wang et al., 2020) ⁸	Princeton University	tool that automatically detects possible forms of bias in a visual dataset

¹<http://www.data-science-public-policy.org/projects/aequitas/>

²<https://aif360.mybluemix.net/>

³<https://aix360.mybluemix.net/>

⁴<https://github.com/pymetrics/audit-ai>

⁵<https://deon.drivendata.org/>

⁶<http://fairlearn.org/>

⁷<https://github.com/linkedin/lift>

⁸<https://github.com/princetonvisualai/revise-tool>

Table 2.1 continued from previous page

Tool	Developer	Features
FAT Forensics ⁹ (Sokol et al., 2020)	University of Bristol	fairness, accountability and transparency python package to evaluate and compare different algorithms
What-If tool ¹⁰ (Wexler et al., 2019)	Google	in hypothetical situations, analyze and visualize model behavior across for different ML fairness metrics

Table 2.1 – Software toolkits for data scientists and practitioners addressing AI Fairness.

⁹<https://fat-forensics.org/>
¹⁰<https://pair-code.github.io/what-if-tool/>

AI Fairness 360 (Bellamy et al., 2018a) by IBM provides a set of tools to implement several machine learning algorithms and data sets commonly used in fairness research. *Fairlearn* by Microsoft is a similar tool. *Aequitas* (Saleiro et al., 2019), an open-source bias audit tool, specializes in comparing several fairness metrics and analyzes how they evaluate different sub-groups. *What-if tool* by Google is focused on comprehensive and customizable visualizations to show individual-level explanations under different scenarios.

Many other individual researchers publish their approaches on GitHub: *Fairml* (Python toolbox that helps audit machine learning models for bias) (Adebayo, 2016), *SHAP* (a game-theoretic approach to explain the output of any machine learning model) (Lundberg & Lee, 2017), and *lime* (used for classifiers that use text, tabular data or images, to explain an individual decision) (Ribeiro et al., 2016). These are practical python libraries for explainability and auditing black-box models. However, they are no longer being supported.

Out of the most common toolkits, most provide an integration with python. *AI Fairness 360*, *Fairlearn*, *FAT Forensics* even offer support for *scikit-learn* (Pedregosa et al., 2011) (a popular package used for machine learning making it easier for practitioners to include it with their existing models).

Several tools offer a web interface to help users navigate and choose fairness metrics. *Aequitas* has a web application with a gradual process and guidance to select a group-level fairness metric. In the next step, the tool produces a comprehensible audit report that is relatively straightforward to explain to non-technical stakeholders. *AI Fairness 360*, *Fairlearn*, *What-if tool* are more valuable for technical specialists with a previous knowledge and understanding of fairness.

2.2.4 Limitations

However, open-source fairness toolkits have their limitations and gaps (Lee & Singh, 2021).

Since fairness does not have a single definition, each toolkit has a different vision of what “fairness” means with limited guidance on which approach may be most appropriate. *AI Fairness 360* and *Fairlearn*, although they seem similar as tools, they frame fairness differently. *AI Fairness 360* focuses on “debiasing” the algorithms while *Fairlearn* focuses on the potential harm, avoiding the term “bias”. *Aequitas* also defines fairness with respect to bias, but its primary intention is not to mitigate but to generate an audit report to flag potentially unfair outcomes.

Some of the tools require an extreme learning curve. Most tools aim at data scientists and users with much experience and practice in the fairness literature and statistics. A user who is not familiar with the relevant literature would need days, if not weeks, to understand the different fairness metrics (group vs individual, demographic parity vs equalized odds) or the available fairness algorithms, given that there is inadequate direction on which techniques to apply and when. Even *What-if*, which uses visualizations to make the tool accessible to a

non-technical audience, is difficult to understand.

Toolkits that offer mitigation techniques, such as *AI Fairness 360* or *Fairlearn*, have limited guidance on what metric to choose and leave the user to explore the differences from the academic papers provided at the end of each session.

Another limitation that prevents these tools from being widely used is the gap between real-life applications and the academic use cases these tools use. The use cases come from popular datasets used in the fairness literature (Mehrabi et al., 2019) such as *UCI Adult Dataset* (Becker, 1994) (also known as the “Census Income” dataset from the 1994 census with data indicating whether the income of a person exceeds \$50K/year), the *German Credit Dataset* (Hofmann, 1994) (also from 1994, containing credit records, which is usually used for exploring gender inequalities on credit-related issues), and the *COMPAS dataset* (Angwin et al., 2016) (from 2013 to 2014, containing records for defendants, demographics, and the COMPAS score). The use cases either use data that are too old or too narrowly focussed on the United States. *Audit-ai* for example, was adapted to meet the requirements of the US employment guidelines to be used as an algorithmic hiring tool, which cannot be generalized and used in other settings.

Another issue is the limited customization. After using *AI Fairness 360* one can notice that the tool is hardcoded to their data, and much extra work is needed to adapt the toolkits to a particular use case (as we found later in the implementation of the notebook). A lot of these practicals tools do not work well between them. For example, *What-If* python package offers compatibility with other explainability tools like *Lime* and *Shap*. However, since tools have different fairness definitions and mitigation approaches, they do not work together.

These tools usually concentrate on more straightforward problems, using mostly binary classification in the examples and demos, and applying group metrics to assess the fairness metrics on protected attributes, like gender or race. However, fairness is more complex in real life. The tools are limited in more complicated regression challenges with various protected attributes that could potentially lead to intersectional discrimination.

2.3 Teaching of tech ethics in university curriculum

Technology has social consequences (Winner, 1980), and therefore, it is crucial that engineers can examine and question the ethical and social implications of their work. Universities have a critical role in meeting these ethical concerns. By integrating human values into computing technology, they can educate scientists and engineers to work for the public good, creating a technology that is in line with human values rather than harming them.

Some universities have started including ethics in their computer science curriculum, and the number of stand-alone ethics courses in data science and machine learning programs is rising lately. However, since these programs are comparatively recent, many institutions do not yet have specialized programs in this domain (especially at the undergraduate level) (Saltz

et al., 2019).

In the 2020 paper *What Do We Teach When We Teach Tech Ethics?* (Fiesler et al., 2020), authors analyzed 115 syllabi from university technology ethics courses. The list, first presented in a Medium article (Fiesler, 2018) and collected through crowdsourcing, currently has more than 290 Tech Ethics courses from different universities around the globe ¹¹.

In 2018, the MIT AI and Ethics group created a set of recommendations to the Department of Computer Science on how to integrate ethics into education (Gilpin, 2018).

Harvard University introduced *Embedded EthiCS*, a collaboration between computer scientists and philosophers to integrate ethics modules into courses across the standard computer science curriculum (College, 2017). The *Embedded EthiCS* approach introduces short ethics modules to computer science courses. Instead of having only one single ethics course, these modules are integrated into the core computer science curriculum.

Several US universities have included ethics, like the *Data Science Ethics* at Yale by professor Elisa Celis. It is open for all students from various departments (data scientists, anthropology, sociology, economics, and other departments) as it does not require coding or technical background. They examine ethical concerns, algorithmic challenges, and policy decisions when solving real-world problems through data science.

EPFL, through the Digital Humanities department, has started to include courses that question the traditional hard science approach to machine learning. The *Critical Data Studies* (Choirat et al., 2020) course is part of the new Université de Lausanne (UNIL) / EPFL teaching offer, which proposes to combine knowledge from Social and Human Sciences and Engineering Sciences to tackle complex themes which require an interdisciplinary methodology. The course aims to highlight the methodological challenges and all the biases that can arise from the study of data considered, a priori, as “objective” and natural.

What all of the above approaches try to combine is interdisciplinarity. The digital ecosystem, the data analysis, the massive data, but also the cultural, ethical, and critical challenges posed by massive data (Iliadis & Russo, 2016) need a combined collaboration across several traditional disciplines.

2.3.1 Limitations

However, recently the pedagogy of AI Ethics for computer science students has received criticism. Authors (Raji et al., 2021) analyzed AI Ethics course syllabi exposing their limitations.

Ethics is commonly viewed only as a specialization, something that someone else does. Although universities have ethics courses in their curriculum, they are primarily stand-alone and unconnected from their computer science education (Fiesler et al., 2020). Earlier research

¹¹updated list can be found at <https://cutt.ly/mmvCWO4>

has revealed that when ethics education is a one-time class, students do not significantly consider the importance of the ethics material to their future profession (Saltz et al., 2019).

Other barriers come from the exact nature of interdisciplinarity. From the *Embedded EthiCS* approach, authors (Grosz et al., 2019) observed several common insecurities. On the one hand, philosophers were worried about their lack of technical knowledge. On the other hand, computer scientists were troubled about their limited experience with ethical issues and, therefore, hesitant to address ethical concerns with their students.

Another obstacle appears from the disciplines' different methodologies and vocabularies. In general, in an engineering course, students are used to problems with a unique correct answer. It is a very different way of thought and often challenging for them when there are probably several acceptable answers to ethical dilemmas (Grosz et al., 2019).

2.4 Human Rights frameworks

Human rights, as opposed to ethics, are often better defined and measurable, as most are defined under international or national law. Authors (McGregor et al., 2019) propose the usage of International human rights law as a framework for algorithmic accountability.

The Universal Declaration of Human Rights (UDHR) (The United Nations, 1948) was adopted by the UN General Assembly in 1948 and is the first legal document to set out the fundamental human rights to be universally protected. It serves as the foundation for all human rights treaties adopted under the auspices of the United Nations and many different national constitutional provisions and domestic laws that are legally binding. While the Universal Declaration itself is not legally binding law, it is considered that it has acquired as customary international law which binds all States (UN Office of the High Commissioner for Human Rights (OHCHR) - Europe Regional Office, 2008).

The UDHR, together with the two covenants - the International Covenant for Civil and Political Rights (The United Nations General Assembly, 1966a), and the International Covenant for Economic, Social and Cultural Rights (The United Nations General Assembly, 1966b) - codified the Universal Declaration into legally binding treaties, that is the so-called "International Bill of Human Rights" (UN Office of the High Commissioner for Human Rights (OHCHR), 1996).

In addition, here are a set of international treaties and national laws that code the contents of the Declaration as legally binding norms. The interpretation of these laws and methodologies to apply judicial and quasi-judicial mechanisms has elaborated them at national and international levels, non-legally binding instruments (declarations, principles, guidelines, code of conduct). Those are called "human rights standards".

Some critically essential treaties range from The Convention on the Elimination of All Forms of Discrimination against Women (CEDAW) (The United Nations, 1988), the International Convention on the Elimination of All Forms of Racial Discrimination (CERD) (The United

Nations, 1966), the Convention on the Rights of Persons with Disabilities (CPRD) (The United Nations, 2006), the Convention on the Rights of the Child (The United Nations, 1989), and others.¹²

States that are signatories to these treaties have obligations to respect, protect and fulfil human rights enshrined under these treaties (Rights, 2018). National constitutions and national law may also provide for human rights, sometimes with more concrete or progressive rights provisions.

The human rights regime provides the clarity and certainty of law. It converts voluntary promises of ethical behaviour into compulsory requirements for compliance with established legislation.

The Office of the United Nations High Commissioner for Human Rights (OHCHR) has identified the urgency to find practical ways to prevent and deal with human rights harms related to the development of digital technologies and their use by corporate, government and non-governmental actors, including individual users.

The UN Human Rights can play a unique role in initiating, hosting and stewarding the development of practical and realistic approaches. It is a UN agency that plays a vital role as a convener as they have more substantial convening power with the governments and civil society, and then business actors.

A typical scenario where a human rights framework works better than an ethics framework is when a company uses a machine learning algorithm to select job applicants. The company trains the model to decide which candidates to hire based on the data and abilities of current or former successful employees. In that case, the algorithm can learn from past biases and continue to propose hiring similar people.

From a human rights standpoint, addressing the discriminatory consequences of the algorithm in that scenario is not an issue of ethical concern but a legal obligation (Raso et al., 2018). The biases can have significant implications for the right to freedom from discrimination (United Nations Human Rights Committee, 1983c), the right to equal pay for equal work (United Nations Human Rights Committee, 1983b), and the rights to freedom of expression and association (United Nations Human Rights Committee, 1983a).

2.4.1 Limitations

Although a Human Rights-Based Approach (HRBA) is pretty well developed at an operational level, a HRBA to data science remains hard to apply.

The challenge of using a Human Rights-Based Approach (HRBA) for designing an algorithm

¹²A list of universal human rights instruments can be found here: <https://www.ohchr.org/EN/ProfessionalInterest/Pages/UniversalHumanRightsInstruments.aspx>

is that it has been understood primarily on legal terms. The guidance notes, good practice case studies, commentary on shared dilemmas and policy recommendations that clarify what certain key concepts/expectations of a Human Rights-Based Approach (HRBA) look like in practice concerning the design, development and use of digital technologies are aimed for relevant stakeholders (States, business, senior business leaders, civil national human rights institutions, academics and other experts) (UN Office of the High Commissioner for Human Rights (OHCHR), 2019).

It has not been translated into general guidance for programmers, computer engineers, and data science practitioners, making it hard for practitioners to apply the guidelines in their everyday tasks.

3 Methodology

We have identified the gap in computer science education that focuses only on science and not the holistic implications coding has for and on society.

In order to address this problem, we asked the following research question:

RQ: How can International Human Rights frameworks be integrated with current concepts of fairness for the design of an educational tool for computer scientists?

To answer this question, we created a methodology that incorporates human rights concepts with a hands-on data science approach.

The approach was first created in 2019 and presented as a theoretical workshop in 2020 at EPFL by the NGO Women at the Table (W@tt) in collaboration with Office of the United Nations High Commissioner for Human Rights. We developed and deployed the applied version of the workshop for the 2021 edition (that is the scope of this thesis).

A Human Rights-Based Approach (HRBA), a core concept of this workshop, is a conceptual framework without a universal definition. Different actors are using slightly different versions of a HRBA depending on their mandates and contexts. The UN common understanding on a Human Rights-Based Approach (HRBA) to development cooperation is defined as “a conceptual framework for the process of human development that is normatively based on international human rights standards and operationally directed to promoting and protecting human rights” (U. N. S. D. Group, n.d.).

The key elements of the UN common understanding of a Human Rights-Based Approach (HRBA) are mainly common to other versions of a HRBA. A HRBA should comply with the human rights law, and the goal should further the realisation of human rights as laid down in the Universal Declaration of Human Rights (UDHR) and other international human rights instruments. These legal instruments should guide all phases of the programming process (UNSDG Human Rights Working Group, 2003).

For designing and developing algorithms, we apply these key elements of Human Rights-Based Approach (HRBA) to the process of decisions made at various points of the data and model lifecycle.

Through the practical Jupyter notebook, our methodology explores how human rights interplay with decisions made at various points of the data and model life cycle. We intend to scale the methodology and workshops by blending workshops at new participating universities with a guest Professor or Doctoral Candidate from the participating university sharing their applied research and articulating how their research interplays with a Human Rights framework.¹

3.1 Overall approach

3.1.1 Goal, learning objectives and outcomes

Goal

The *<AI & Equality> A Human Rights Toolbox* workshop is designed to demonstrate to computer science students how to apply a Human Rights-Based Approach (HRBA) to AI.

Learning objectives

Applying a Human Rights-Based Approach, we use gender as a proxy for learning about human rights discriminations; equally applicable for race, disability, age, or any other form of discrimination. The workshop aims to develop and strengthen awareness and understanding of gender equality and gender bias as the first step towards behavioural change. It aims to integrate an intersectional perspective into the everyday work of computer science and engineering. Intersectionality, as defined by civil rights activist and professor Kimberlé Crenshaw is “the interconnected nature of social categorisations such as race, class and gender as they apply to a given individual or group, regarded as giving overlapping and interdependent systems of discrimination and disadvantage” (Crenshaw, 1989).

Throughout the workshop, participants complete a variety of interactive exercises, discussions and activities. Specific training materials support the workshop, the main one being the Jupyter notebook, along with supportive slides and extra material tailored for students to embed gender and equality across their research and day-to-day work.

The workshop showcases a HRBA to AI, provides examples of how gender, racial and other biases can occur in the research, design and development of algorithms and illustrate methods, tools and techniques to mitigate bias in AI.

¹The material presented at the workshop can be found at <https://aiequalitytoolbox.com/resources.html>

Learning outcomes

By the end of the workshop, the participants must be able to:

1. Explain a human rights-based approach to AI
2. Identify the relevance of different biases and importance of intersectionality, gender equality and bias to computer science and engineering / institutional objectives
3. Analyse how gender, racial and other bias has occurred or can occur in the research, design and development of AI
4. Apply how and when to use tools and techniques to mitigate bias in AI
5. Evaluate methods to integrate non-discrimination into design, planning and implementation of AI projects

3.1.2 Beneficiaries

The workshop targets senior undergraduates, graduate students and PhD university students with some experience in Data Science. The participants should have some experience with Python and the Jupyter Notebook, as the practical part of the workshop is developed using Jupyter Notebook.

The participants will benefit the most if they have already applied some data analysis and machine learning algorithms in their work so that they are familiar with the data pipeline (data extraction, analysis and cleaning) and the machine learning pipeline (feature extraction, model training, model evaluation).

3.1.3 Workshop structure

We separate the workshop and the material into two main thematics, theory and practice, to facilitate the learning.

We employ a broad to narrow approach, going from theory to applied practices. In the first part, we give an introduction to basic human rights principles. Then we explore how this plays out in current research. In the second part, we present practical ways to translate human rights principles through code and ways to identify and mitigate algorithmic bias.

A different expert in the specific domain presents each part. At the end of each session, we strongly encourage discussion, questions, and knowledge sharing with and between the students.

The structure, analysed in the following sections, is:

Part I A. Human Rights Module Taught by human rights / legal experts introducing basic human rights concepts and a Human Rights-Based Approach (HRBA) to machine learning

B. Applied Research Research Representatives (PhD students, post-doc, faculty) present their work on how human rights fit with AI

Part II A. Practical Toolbox Step-by-step case study, an interactive way to see how to apply Human Rights-Based Approach (HRBA) in practice (debiasing data and algorithms)

3.2 Human Rights Module

3.2.1 Objectives

The objective of this session is to understand key concepts of international human rights law which are relevant to the designing of algorithms, with a focus on principles of equality and non-discrimination.

Learning outcomes

In this module, students are expected to:

- be able to describe key concepts of equality and non-discrimination under international law
- be able to explain why social norms and stereotypes can cause discrimination and how this is relevant to the function of algorithms
- be able to describe key elements of a human rights-based approach relevant to algorithms

3.2.2 Structure

A human rights or a legal expert teaches this part. The Human Rights modules from the first three workshops were designed and led by a dedicated Human Rights Officer from OHCHR's Women's Rights and Gender Section, Asako Hattori, who collaborated on the material and methodology with the instigator of the workshops, Caitlin Kraft-Buchman the CEO/Founder of the Swiss NGO Women at the Table, and EPFL.

The expert presents with slides and backup materials, including a glossary of terms. The session lasts for 20-minutes and focuses on digestible definitions of human rights and human rights principles, equality and non-discrimination. It first introduces a Human Rights-Based Approach (HRBA) and poses a critical analysis of what this approach might look like in machine learning. After the presentation, a 10-15 minute discussion follows.

Roadmap

The flow of information goes from the concept to the application and then to the resources.

- Understanding key concepts
 - What are human rights?
 - Human rights principles
 - Equality and non-discrimination
 - * Equality
 - * Non-discrimination
 - * Social norms and stereotypes
 - * Temporary special measures
- Application of human rights
 - What is a human rights-based approach?
- Legal sources
 - Where can we find the legal basis of human rights?

3.2.3 Content

The presentation starts building the foundation: “what are human rights?”.

Since the participants are from a technical background, they may have never heard these definitions before.

According to the Office of the United Nations High Commissioner for Human Rights (OHCHR)

“Human rights are rights we have simply because we exist as human beings. These universal rights are inherent to us all, regardless of nationality, sex, national or ethnic origin, color, religion, language, or any other status.”

(The United Nations, 1948) (Rights, 2018)

Human rights are universal, indivisible, inter-dependent and inter-related (Rights, 2018). Universality means that human rights, such as those defined in the Universal Declaration of Human Rights (UDHR) should be guaranteed to everyone everywhere in the world. This, in turn, means that all of us are also responsible for respecting other people’s rights. They are rights we have because we exist as human beings - they are not granted by any state.

Indivisibility and inter-relatedness mean that one set of rights cannot be enjoyed thoroughly without the other.

For example, in 2019 in New York, claims were made that the algorithm used by a credit company for evaluating credit limits of customers granted lower credit limits to women than men (Hansson, 2019) (Nasiripour & Farrell, 2021). The long-term structural gender inequality, equality in access to financial resources (which is a recognised human right), is linked with equality in the enjoyment of the rights to education, to work, at work, equality in family life.

The workshop is focused on selected Human Rights principles under the UN Common Understanding on a Human Rights-Based Approach (HRBA) (The United Nations, 1948)²:

Equality and Non-discrimination All individuals are equal as human beings and by the inherent dignity of each human person.

Participation and Inclusion Every person and all peoples are entitled to active, free and meaningful participation in, contribution to, and enjoyment of civil, economic, social, cultural and political development in which human rights and fundamental freedoms can be realised.

Accountability and Rule of Law States and other duty-bearers are answerable for the observance of human rights.

One important note and misunderstanding, is about the terms *equality* and *equity*. In the algorithmic fairness literature, *equality* is defined as treating everyone the same way, and *equity* means treating each individual, taking into account the different situations of each one. However, under the international human rights law, equality means “equal enjoyment of human rights” (The United Nations, 1948). Human rights frameworks not only “guarantee” everyone the same treatment but also guarantee that everyone can enjoy their rights equally. This principle of equality under international human rights law requires specific efforts to ensure de facto equality or substantive equality (Figure 3.1). For example, to reverse the perpetuation of historical and institutionalised discrimination, active use of temporary special measures, sometimes also called “affirmative actions”, may be required (of the United Nations High Commissioner for Human Rights, 2004). The concept of equality and equity are complementary, and both discourses should inform each other.

The following key concepts that are introduced are discrimination, social norms and stereotypes and temporary special measures.

The international bill of Human Rights prohibits discrimination based on: “race, colour, sex, language, religion, political or other opinions, national or social origin, property, birth or

²Other key principles are *Universality* and *Inalienability*, *Indivisibility*, *Inter-dependence* and *Inter-relatedness*. The expert focuses on the selected principles to make the introduction to human rights more simple and concrete.

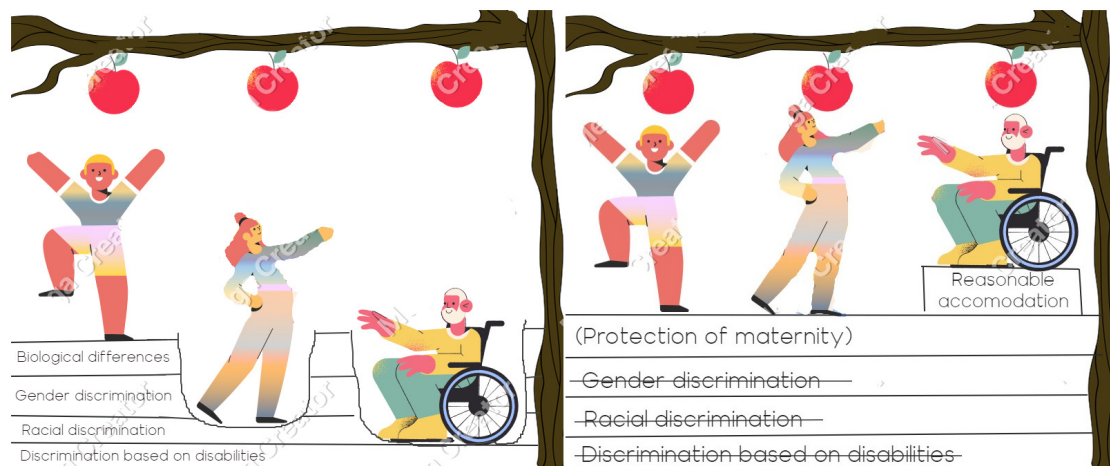


Figure 3.1 – Equality under international human rights law. Equality means "equal enjoyment of human rights". It's not only about guaranteeing everyone the same treatment, but also about guaranteeing that everyone can enjoy their rights equally. (Asako Hattori, 2020)

other status" (Article 2 of each instrument) The prohibited grounds of discrimination under international human rights law have been further elaborated over the years,

After this, the presenter introduces questions about the use and the role of algorithms in discrimination. For example "Should we make our tool 'blind' to these characteristics to avoid discrimination?" "Do we have to worry about discriminatory 'impact' of algorithms?" "How do we approach a combination of several types of discrimination (intersecting forms of discrimination)?"

These questions aid with the discussion at the end of the presentation so that students start using critical analysis regarding their tools and code. They begin to view the machine learning system as a whole entity.

A Human Rights-based approach

Although there is no universal definition of Human Rights-Based Approach (HRBA), there are some common structures and elements.

A Human Rights-Based Approach (HRBA) should:

1. Comply with the human rights law
2. Goals should contribute to the realisation of human rights
3. Processes are guided by human rights principles

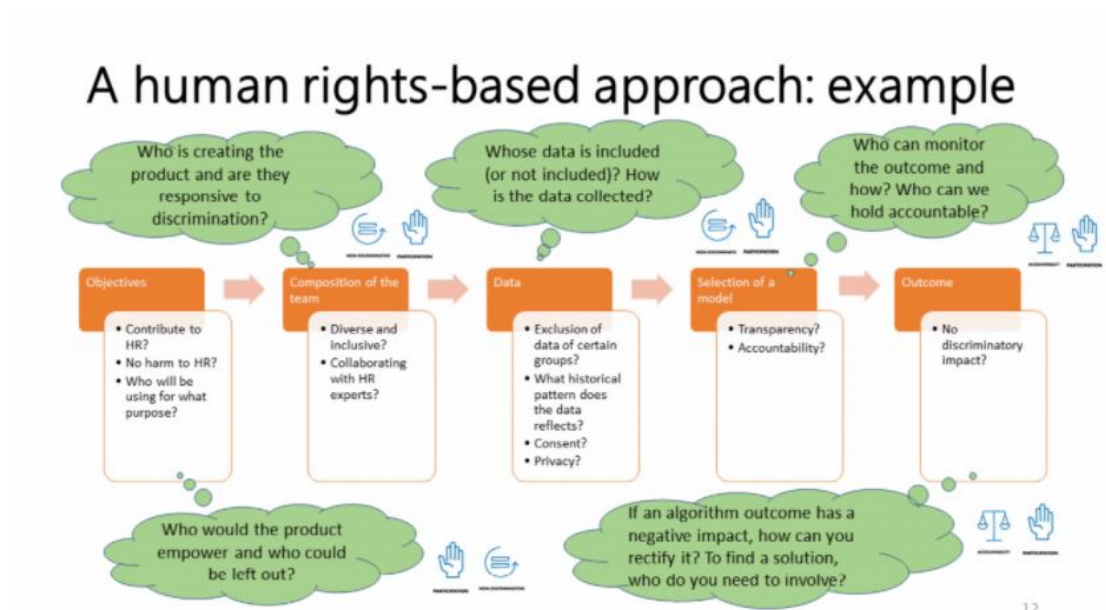


Figure 3.2 – A Human Rights-based approach: an example of key questions relevant to human rights principles to ask at each stage of a machine learning algorithm development. (Asako Hattori, 2020)

4. Empower right-holders³ and duty-bearers⁴

Figure 3.2 shows some of the critical questions that should be asked during the process that aligns with the principles of non-discrimination, participation and equality. Participants start to see the development of the algorithm as a whole process, from the team's composition to the objectives that will create the algorithm, the data that will be used, the model selection and the outcome of the process. These stages are explained in the following parts of the workshop as well.

3.3 Applied Research

3.3.1 Objectives

The objective of this session is to review the social impact of algorithms. The scope is for participants to connect human rights and current research and for the presenters to showcase their work linked between AI and human rights, the social impact of machine learning algorithms or similar topics.

³Individuals and groups (e.g. indigenous people as collective) with valid human rights claims

⁴States and non-state actors (incl. business enterprises) with corresponding obligations towards human rights.

Learning outcomes

In this module, depending on the research presented, participants are expected to:

- situate the impact of inequalities produced by development choices through a practical case study
- reflect on the relevance of critical thinking and the justification of choices for better development practices
- stay aware of the social and human implications of the data science and engineering profession
- discuss collectively how a human rights-based approach could help implement better algorithms

3.3.2 Structure

PhD students, post-docs or professors teach this part, which lasts 20 minutes, followed by a discussion.

In this session, participants see practical examples of how these concepts interrelate. They also become aware of research and collaboration opportunities at their university. The researchers can increase the visibility of their work and potentially find new collaborators either within the university or the larger AI Equality community.

At EPFL, this section was presented by the PhD candidate, Jessica Pidoux, who talked about her thesis on online dating algorithms (Pidoux, 2019) (Pidoux et al., 2021). She explores the impact of inequalities that have been re/produced at scale either intentionally or with unintended consequences due to the choices made at the development and deployment stages of the Tinder algorithm.

Going forward to scale, we propose that one researcher from the participating university hosting the workshop present their applied research as it relates to the theme of human rights and machine learning linking human rights concepts laid out in Part I with current research on AI, human rights, the social impact of machine learning algorithms or similar topics.

3.4 Practicum

3.4.1 Objectives

The objective of this session is to experiment with data to see how different mathematical and data concepts of fairness interrelate, begin a critical analysis checklist of the data process and apply some of the concepts and debiasing literature to hands-on exercise.

Learning outcomes

In this module, students are expected to:

- introduce fairness metrics and apply the baseline model
- explore where bias can be found in the machine learning pipeline
- be able to understand where we can intervene in a fairness pipeline
- question the data and the model choices

3.4.2 Structure

This part is developed and taught by the author, and it lasts 30 minutes. It is separated into three time slots, of 30 minutes each: 15-minute presentation, 5-minute practical work, followed by a 10-minute discussion between the participants and the presenters and then a 5-minute break. This structure best fitted the time constraints presented by the organisers. However, different teaching strategies can be tested in the future versions of the workshop.

Time schedule:

1. Introduction to fairness metrics & baseline model
2. Bias in data: pre-processing
3. Bias in model and outcomes: in-processing and post-processing

In terms of usability, slides summarising the main concepts, as seen with figure 3.3, were created.

This clear methodology provokes critical analysis on where and when to intervene in the fairness pipeline: pre-processing (training data), in-processing (model design) and post-processing (predictions).

Throughout the workshop, we used the German Credit Dataset (Hofmann, 1994), commonly used in algorithmic fairness literature (Verma & Rubin, 2018), as a case study. The goal is to predict whether a person can repay a loan. We explore how this can be done with respect to a Human Rights-Based Approach. We also show different sources of bias and ways to mitigate it.

This presentation (with annotations and links for deeper analysis before the workshop or further work in the notebook) is followed by a discussion.

3.4.3 Roadmap

- Introduce fairness metrics

- Perform exploratory data analysis and create a baseline model
- (Pre-processing) Rebalance the data
- (In-processing) Build a model with fairness constraints using a Meta Classifier
- (Post-processing) Optimise for the different groups of people using Equalised Odds

3.4.4 Introduction to fairness

First, we introduce students to what is fairness and the different definitions and examples of how different definitions of fairness work out.

Objectives

The objective is to understand that fairness is complex:

- it is not technical, but an ethical concept
- it is contextual, as there is no one-size-fits-all approach
- there are no set answers and often cost, and benefit decisions have to be made
- it is a process, and there is no single fairness checkpoint

Content

What is fairness?

Fairness in machine learning has captured the attention of researchers in the AI, Software Engineering, and Law communities, with a steady increase in related work over the past few years. Many efforts have been made to “debias” the data and create a “fair” model. But what is fairness? And how can this translate to machine learning?

Starting with these questions, we clarify: Fairness is not a technical or statistical concept, and there can never be a tool or software that can fully “debias” your data or make your model “fair”. Fairness is an ethical concept and a contested one at that. At best, we can select some ideal of what it means to be “fair” and then make progress toward satisfying it in our particular setting (Khan et al., 2021).

We can then start examining data science and research through the lens of power. The question “fair to whom?” lies at the centre of the data, the model, and the system. Fair *to whom*, for *whom*, *by whom*, and with *whose interests and goals* are questions core to a basic and ethical design process (D’Ignazio & Klein, 2020).

Answering these questions, we see there are many different groups, stakeholders, and different interests and goals, leading us to many definitions. There are different schools of thought for fairness and ethical decision making, each of them containing several varieties of approaches to ethics (Bonde & Firenze, 2013). There are 21 mathematical definitions of fairness at last count (Narayanan, 2018).

The notion of fairness is contextual in both societies and mathematics (Binns, 2021). The paper *Fairness definitions explained* (Verma & Rubin, 2018) takes 20 of these definitions and applies them to the German Credit dataset that we are using for these exercises.

So as there is no one definition of fairness or ethics, there is also no explicit consensus on which definition to use in each case. These definitions can even cancel one another out in what is called the impossibility theorem (Kleinberg et al., 2016). The theorem states that no more than one of the three fairness metrics (risk assignments) of demographic parity, predictive parity, and equalised odds can hold simultaneously for a well-calibrated classifier, and we need to make trade-offs.

But why are there so many definitions? Why and how do we consider the same case fair using some definitions of fairness but unfair if we use others?

To help with learning, we show how three different fairness metrics, two for group fairness and one for individual, differ conceptually, where they work, and what gaps they create.

In the presentation, we use job applications between male and female candidates to see how these concepts play out.

Another concept that we use in the examples and is standard in the fairness literature, is the term *protected attribute*⁵, which is an attribute recording sensitive information about individuals. In most cases, the law identifies some characteristics on which it is illegal to discriminate. These traits are usually considered “protected” or “sensitive” attributes in computer science literature (Chen et al., 2019). A protected feature should not be used in an algorithm to determine the outcome for an individual (Barocas et al., 2019). Fairness is often monitored for groups of people protected by anti-discrimination laws, such as subgroups defined by gender, ethnicity, age, disability status. (Mehrabi et al., 2019).

Group fairness metrics

One way to define fairness is using group fairness metrics (or group statistical property), where “groups should receive similar treatments or outcomes”. In our example, this meaning groups of people grouped by gender like males or females should receive similar job acceptance rates.

Two popular definitions are *Demographic Parity* (Zafar, Valera, Rodriguez, et al., 2017) and

⁵We use *attribute* for the quantity describing an instance (ex. gender) and *feature* for is the specification of an attribute and its value (ex gender is female).

Equalized Odds (Hardt et al., 2016). We chose these metrics because they are commonly used in the fairness literature, and they have the property of cancelling each other out at the impossibility theorem (Kleinberg et al., 2016).

For the group fairness metrics, we use the following mathematical notation:

- $X \in R^d$: the features of each candidate (level of education, high school, previous work, total experience, and so on)
- $A \in \{0; 1\}$: a binary indicator of the sensitive attribute. Here is 1 for females, 0 for males
- $C = c(X, A) \in \{0; 1\}$: the classifier output (0 for rejected, 1 for accepted)
- $Y \in \{0; 1\}$: the target variable. Here, it is whether the candidate should be selected or not.

Demographic Parity

Demographic parity

The acceptance rates of the applicants from the two groups must be equal.

Group fairness metric

(+)

- $\frac{4}{5}$ rule
- independence of protected variable

(-)

- random guessing for the minority group still satisfies the metric

Comment:

- should have an action plan to fully support the minority group




Figure 3.3 – Example slide from the workshop about demographic parity: positives, limitations, uses

We say that a classifier C satisfies demographic parity if the outcome is independent of the protected attribute A (gender). To translate it into probabilities, it needs:

$$P\{Y|A=0\} = P\{Y|A=1\}$$

To satisfy demographic parity, applicants from the two groups should be accepted at an equal rate (for example, 50% of male and 50% of female applicants get the job).

There are several motivations to use this metric. First, there is legal support behind this rationale - called the four-fifth rule (Dynamics, 2009). Meaning that, in our use case, females who apply and are hired should be no fewer than four-fifths of the males who get the job.

Another reason to use this metric is the immediate and long term benefits. The intervention rebalances the numbers (in this case) of female and male hires. This intervention may have long term benefits as well (Hu & Chen, 2018).

However, it does not guarantee fairness. First, this metric allows a classifier to pick qualified candidates for the demographic group $A = 0$ (the majority group), but unqualified candidates for the group where $A = 1$, provided that the acceptance percentages match. This situation can typically happen when we do not have enough training data for the minority group. As a consequence, the hiring company might have a deeper understanding of whom to hire in the majority group while random guessing within the minority (Hardt et al., 2016).

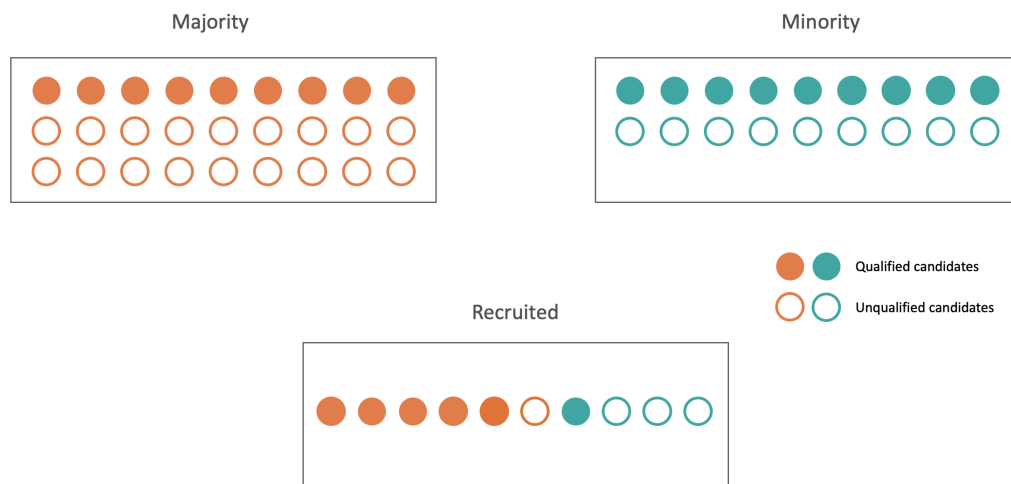


Figure 3.4 – Demographic parity in job applications with laziness: Nothing would prevent the use of a trained model to select candidates from the majority group, while candidates from the minority group were selected randomly with a coin toss — as long as the number of selected candidates from each group is valid. (Landeau, n.d.)

There are cases, though, that demographic parity should be used. For example, we can use it if we know the historical bias that may impact our data. We need to develop a strategy to support the historically marginalised group. Recently, Oxford University aims to promote diversity by accepting more students from underprivileged backgrounds. Their plan also includes extra support for students from disadvantaged backgrounds before beginning their degree courses (Coughlan, 2019).

Equalized odds

Equalized odds (Hardt et al., 2016) also called Separation, Positive Rate Parity (Zafar, Valera,

Rodriguez, et al., 2017), requires that the positive outcome is independent of the protected attribute A , conditional on the actual Y .

The mathematical formulation is:

$$P\{\hat{Y} = 1|A = 0, Y = y\} = P\{\hat{Y} = 1|A = 1, Y = y\}, y \in \{0, 1\}$$

It translates as the probability of hiring a qualified applicant, and the probability of not hiring an unqualified applicant is the same between male and female candidates.

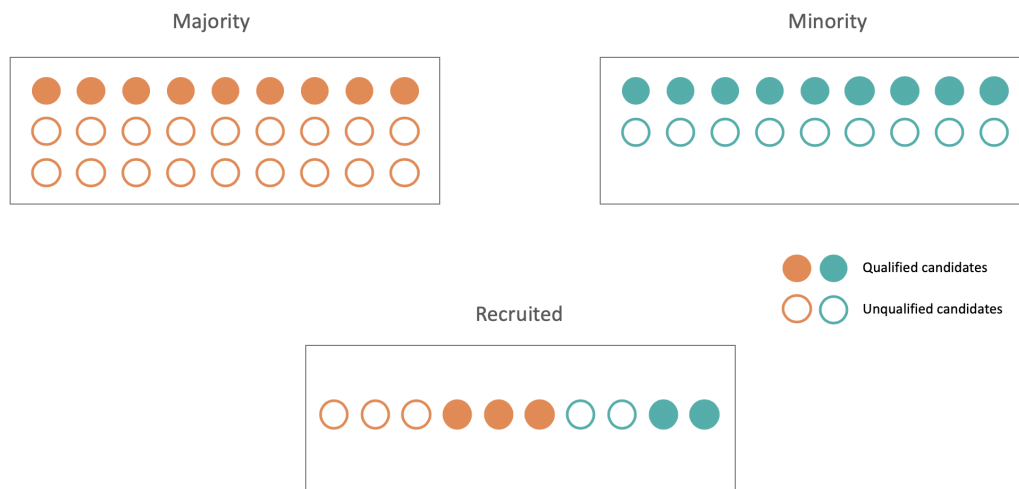


Figure 3.5 – Equalized odds in job applications: Among all the recruited candidates, the same proportion of really qualified applicants should be the same (Landeau, n.d.).

It does not take into account unfairness preexisting among candidates and replicates it. Models corrected under predictive rate parity can also boost unfairness in the long term. In practice, this method also has definition flaws: it needs access to the actual value of the target, which is sometimes hard to define (actual qualification for a job as an example). It is very similar to equality of opportunity but way more challenging to implement in practice.

The motivation to use this metric is that, compared to the demographic parity example, if many unqualified male applicants apply for the job, the hiring of qualified female applicants in other protected groups is not affected. Unlike Demographic Parity, where we hire 50-50 randomly, which counts as fairness, this definition selects between appropriate people from both groups to hire. It, therefore, penalises laziness because hiring unqualified applicants get penalised.

However, it still does not ensure fairness. First, an equal odds classifier classifies everyone of the same low-quality result as the hardest group, which is why it costs over twice as much in this case. Using this metric also leads to a more conservative allocation of job positions, so it

is slightly harder for job-fitting people of all groups to get the job (Hardt et al., 2016).

Secondly, it still might not help close the gap between the two groups in the long term. In the presentation, we use the following example (data from table 3.1): We assume that a company has 30 open positions for the same job. In our scenario, we have two groups with 100 candidates each. In group A, 58 of them fit the job description, while only 2 of them qualify for this job in group B.

If the company decides to hire using the equality of odds metric, it will accept 29 candidates from group A and only one from group B. If this is a well-paid job, then consequently people from group A can afford a higher living standard, better education for their children that, as a result, it will allow them to be qualified for these jobs when they grow up. Over time, the difference between the two groups will increase.

We can choose the equalised odds fairness metric when we want to have a correct prediction on the positive outcome while we care to minimise the false positives. The loan application is an excellent example to use this metric because we care to correctly identify applicants that can repay the loan, as it would increase the profit for the bank. We also want to reduce giving a loan to people who cannot repay it.

Individual fairness metrics

Another way to define fairness metrics is from the perspective of the individuals. In the individual fairness metrics, “similar people should be treated similarly”. We will use the Generalised entropy index (Shorrocks, 1980), which is a measure of inequality (used in economics to measure the distribution of income and economic inequality).

Generalized entropy index

This index measures how evenly members of gender groups are distributed within the application we want to create. If the index is zero, then we have absolute equality. The higher values signify higher levels of inequality.

The motivation behind using individual fairness metrics is that they are more fine-grained than any group definitions of fairness: it restricts the treatment per set of individuals. It has been proposed as a measure of income inequality in a population.

Individual fairness is based entirely on the definition of “similarity” between applicants. That can create new fairness problems if this metric misses critical information. It is not easy to define a suitable metric function to measure the similarity between two inputs.

In a simpler set of job applications, we assume we have three candidates for the same job. The first has a bachelor’s degree and professional experience on the job, the second has a master’s degree and the same professional experience, while the third also has a master’s degree but no relevant experience. Then how do we define the similarity between these applicants? How

	Qualified	Not qualified	Total
Group A	58	42	100
Group B	2	98	100
Total	60	140	

Table 3.1 – Job application example: A company has 30 open positions and candidates are separated by groups (A & B), and by who fits the job description

closer is the first applicant to the second or the third? It can be more complicated if we choose to consider sensitive attributes to this function.

Trade-offs

Groups Fairness vs Individual Fairness

Fairness metrics typically emphasise either on individual or group fairness. However, they are usually unsuccessful in combining both. When it comes to group fairness, most approaches deal with between-group issues (for example, between groups of different gender or race), but this can increase the within-group unfairness (between members of the same group). Decreasing the between-group unfairness can worsen the individual unfairness, which causes an increase in the overall unfairness (Speicher et al., 2018).

Participants see this trade-off in practice later in the practical session.

Demographic Parity vs Equalized Odds

Even between the group fairness metrics, we cannot satisfy both Demographic Parity and Equalised odds.

The example below illustrates this impossibility. In our example, a company has 30 open positions, and on table 3.1 we see the candidates separated into groups (ex. gender: males/females), and then separating the groups by who fits the job description.

Using demographic parity, first, we look into these two groups and out of these, we select half from one group half from the other. In the end, the company will hire 15 applicants from both groups (so inevitably hiring some unqualified applicants).

Using equalised odds, first, we look into the people that fit the job description. That way, the company will hire 29 applicants from group A and one from group B. It is mathematically proven that either Demographic Parity holds or Equalised Odds but not both.

Fairness versus Accuracy Trade-off

Choosing what is fair comes with a cost. Creating a general quantified notion of fairness is extremely difficult. (Wick et al., 2019) (Rodolfa et al., 2021). These quantitative definitions

reduce fairness to one more performance metric in the evaluation of an algorithm. However, improving fairness often leads to lower overall accuracy in the model. It is necessary to analyse the potential trade-offs in a given scenario. Sometimes, greater accuracy in a model can lead to more significant unfairness.

Most technical students are trained for accuracy, optimising for correct predictions, but why not think about how it affects people? Trying to optimise for accuracy can potentially minimise fairness and cause harm.

So we want to be fair to whom? Furthermore, what happens with intersectionality (Foulds et al., 2019)? An algorithm that is fair in terms of gender and race could be unfair in the intersection (for example, women of colour) (Buolamwini & Gebru, 2018).

Conclusion

In conclusion, the participants start to see that the discussion about fairness is not an easy one. First, fairness is highly contextual. There is no one-size-fits-all approach, and it depends on the stakeholders and the application.

Second, we see that there are no set answers. Many times, cost and benefit decisions have to be made.

Finally, fairness is not a “measurement” as it implies a straightforward process but a continuous process. It should be seen as an investigative process that requires detection, explanation and mitigation. There is no single fairness checkpoint; harmful properties can enter a system under biased data or through data science practices and decisions. This triggers the need for internal solid governance, checklists and monitoring.

3.4.5 Baseline Model

It is the first time the participants use the Jupyter notebook and interact with the code. We create a simple baseline model, without applying any techniques, to use as a reference when we later apply the debiasing techniques and perform an exploratory data analysis. The code for the model and the analysis already exists; however, interested participants can explore further and try their approaches.

The instructor guides this part, runs parts of the code and explains the steps and logic, helping to begin familiarising with good data practices that interrogate the dataset’s origin, motivation, and content.

For this and the following sections, we ask questions mentioned in the previous parts.

The goal is for students to start questioning what has traditionally (as part of their education) been taken for granted: the neutrality of the data or the inevitability of working with incomplete

or biased data.

Objectives

In this section, participants are expected to:

- start questioning the data process
- explore the data
- create a baseline model
- see how it performs on the above fairness metrics

Content

We start with a series of questions before dealing with the data and then dig deeper with an example of what we mean by exploratory data analysis.

We employ the German Credit Dataset (Hofmann, 1994), commonly used in fairness literature. We chose it because different fairness algorithms and metrics are implemented in the python package of the aif360 library (Bellamy et al., 2018a).

For this and the following sessions, we have curated a list of questions taken from the work *Datasheets for Datasets* (Gebru et al., 2018) and *Model Cards for Model Reporting* (M. Mitchell et al., 2018).

Biases can be introduced into an AI system at any stage of the development process, from the data collected to the way it is collected, which algorithms we use, what assumptions we make. By asking the right questions upfront during application design, we can prevent many of them.

So we start by asking about the origin and the motivation of our data (Gebru et al., 2018):

- “Why was the dataset created?”
- “For what purpose was the dataset created? Are there tasks for which the dataset should not be used?”
- “Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organisation). Who funded the creation of the dataset?”

These questions help identify early different types of biases.

For example, there could be tasks for which the dataset should or should not be used. The dataset *Labeled Faces in the Wild* (Huang et al., 2008) was created to provide images for face

recognition tasks in an unconstrained setting where image characteristics, subject demographics, or appearance cannot be controlled. We could use the dataset for a face identification problem but not for tasks where the stakes are high (ex. law enforcement).

Also, by being transparent and clear about funding interests, we can identify and mitigate a source of bias called funding bias: “We have funding bias when biased results are reported in order to support or satisfy the funding agency or financial supporter of the research study” (Krimsky, 2013).

Then the participants load the dataset. The initial dataset has 1000 entries with 20 categorical/symbolic features. Each entry in the dataset describes a person who applies for a loan from a bank. For each observation, there are 20 variables, including demographic information (e.g. age, gender) and financial information (e.g. savings and credit history), and a decision outcome representing whether the applicant has a good or bad credit risk.

The protected attributes are “sex” (“male” is privileged and “female” is unprivileged). The outcome variable is “credit-risk”: good (favourable) or bad (unfavourable) (Kamiran & Calders, 2009).

Exploratory Data Analysis

Throughout the exploratory data analysis part, we combine the human rights principles with the code practices through the questions and coding.

For example, the human rights principle of participation and inclusion is brought to the fore through the following question (Gebu et al., 2018)

“Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset”

Starting from this question, students explore the demographic characteristics of the dataset. The questions to guide the analysis are: “Who is in the data?” “Who is missing?” “What is being overgeneralised?” “What is being underrepresented?” “How do the variables, and their values, reflect the real world?” “What might we be leaving out?”

From the generated plots (Figure 3.6), students already see that males overrepresent the females in the dataset with a ratio of almost 7:3. The data is very imbalanced, and if we do not take any action, it will affect our model.

It is also important to see the target variable for gender (Figure 3.7). From these plots, we see that the ratio $\frac{\text{good credit score}}{\text{bad credit score}}$ is bigger in males than females.

Without pre-processing the data, the algorithm would probably show a flavour towards male applicants.

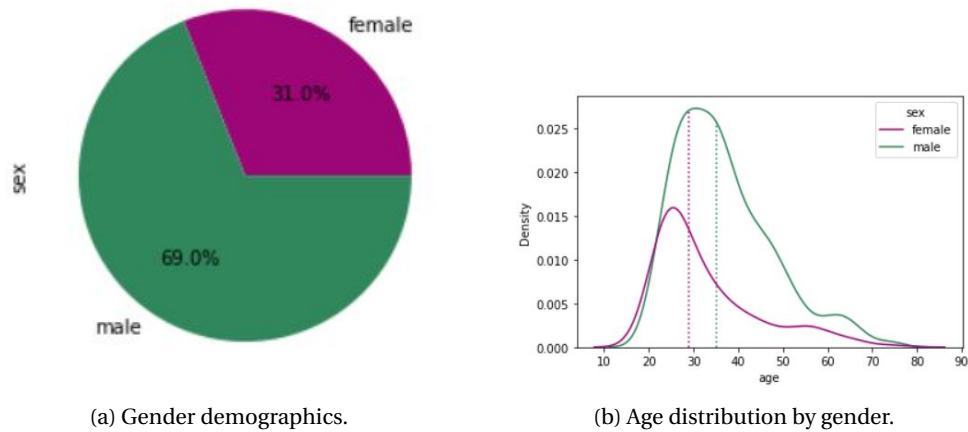


Figure 3.6 – Exploratory demographics plots: Males overrepresent the females in the dataset with ratio almost 7:3. In general, women are younger when they ask for a loan, compared to men.

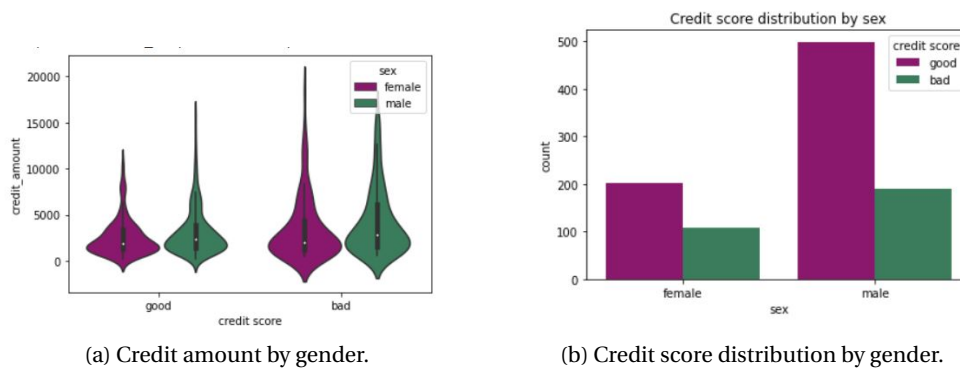


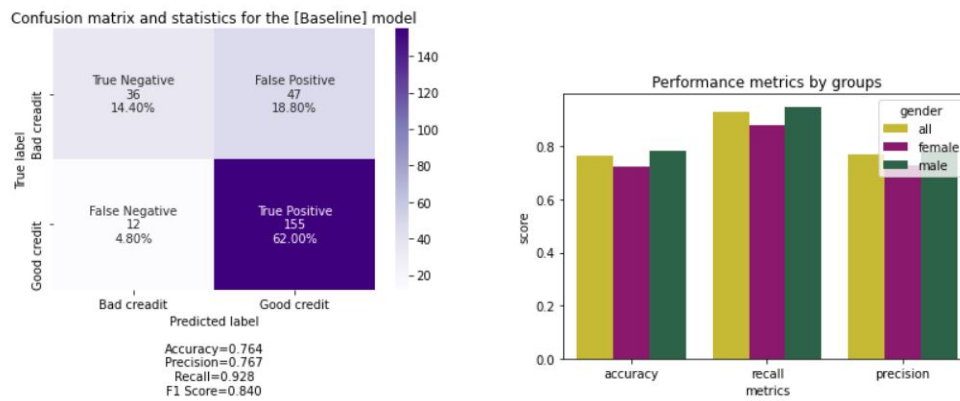
Figure 3.7 – Exploratory target plots: Females who ask for a bigger loan amount are more likely to receive a bad credit score than males.

These plots (Figures 3.6 and 3.7) are essential to a human rights-based approach, as participants start to understand the inequalities between the groups. This analysis is just an introduction. The analysis can go deeper by analysing more intersectional groups, like young men and women vs older men and women, or adding the foreign work status.

The take-away from this analysis is that the more we know our data, the more we can understand and know when, where and how to intervene to achieve an outcome concerning human rights.

Fairness metrics

We then apply the fairness metrics in the baseline model to explore how the model performs in terms of fairness. We highlight the importance of evaluating the model for each subpopulation



(a) Confusion matrix for the Baseline model.

(b) Performance metrics for the Baseline model for each group.

Figure 3.8 – Performance metrics for the algorithm, and then evaluation metrics broken down by gender.

(in our example, gender: male and female loan applicants).

When we evaluate the overall performance of a model, we get some insights into its quality. However, it does not give us much information about the performance of the model in different subgroups.

When we want to evaluate a model for fairness, it is essential to see if the prediction errors are the same between the different subgroups. It is crucial to identify if a specific group of people is more vulnerable to prediction errors than others.

One way to compare and choose classifiers is accuracy. We define *accuracy* for a model as the number of correct predictions divided by the total number of predictions. It is easy to calculate and interpret and summarises in one single number the model's capability. However, it misses some other critical aspects.

A more detailed approach is to look at the *confusion matrix*. Most data science and machine learning practitioners are familiar with its use to evaluate a model. A *confusion matrix* is a grid that plots predictions vs ground truth for the model and tabulates statistics summarising how often the model made the correct prediction and how often it made the wrong prediction (Powers, 2008).

The plots (Figure 3.8) show that using default parameters, we find that the model performs better for males than females. Specifically, we find that accuracy, recall (proportion of actual positives was identified correctly), and precision (proportion of positive identifications was correct) for females are worse than males.

One of the reasons could be the rate between the genders. We noticed that males are disproportionately more than females in the data set (more than 2-to-1).

Through this confusion matrix demonstration, participants find that the results vary slightly from the overall performance metrics, highlighting the importance of evaluating model performance across subgroups rather than in aggregate.

From this evaluation, participants learn that they need to ensure that they make an informed decision regarding the trade-offs among the false positives, false negatives, true positives, and true negatives in their work.

3.4.6 Pre-processing

In this section, we investigate where bias can exist in the data, the different types of data biases, and we use the questions from *Datasheets for Datasets* (Gebru et al., 2018).

We discuss techniques on how to identify potential sources of this bias. Finally, we encourage students to try a pre-processing algorithm.

Objectives

For this section, participants are expected to:

- identify different types of biases that can be found in the data
- try pre-processing techniques to debias the data

Content

Pre-processing is about assessing the training data. The training data is the basis of every predictive model. Data used to measure and categorise people run the risk of causing undesirable properties if they are the base of a model used to support decision-making.

Bias in the Data

Bias in data can exist in many shapes and forms, leading to unfairness in various learning tasks.

Many data science practitioners have been told that “data doesn’t lie”, or that “data beats opinion” (Lehner, 2020).

By showing the participants the different types of biases introduced through the data collection process, they start to question the objectivity of the data.

However, since data reflects a part of the society - with the systemic biases that come with it - and it is collected and created by humans - with the statistical biases (Figure 3.9) that come with it, we can conclude what academic Kate Crawford notices: “Data sets are not objective; they are creations of human design.” (Crawford, 2013) and as data journalist Lena V. Groeger

wrote in her ProPublica article *When the Designer Shows Up In the Design* writes “Data does not speak for itself – it echoes its collectors” (V. Groeger, 2017).

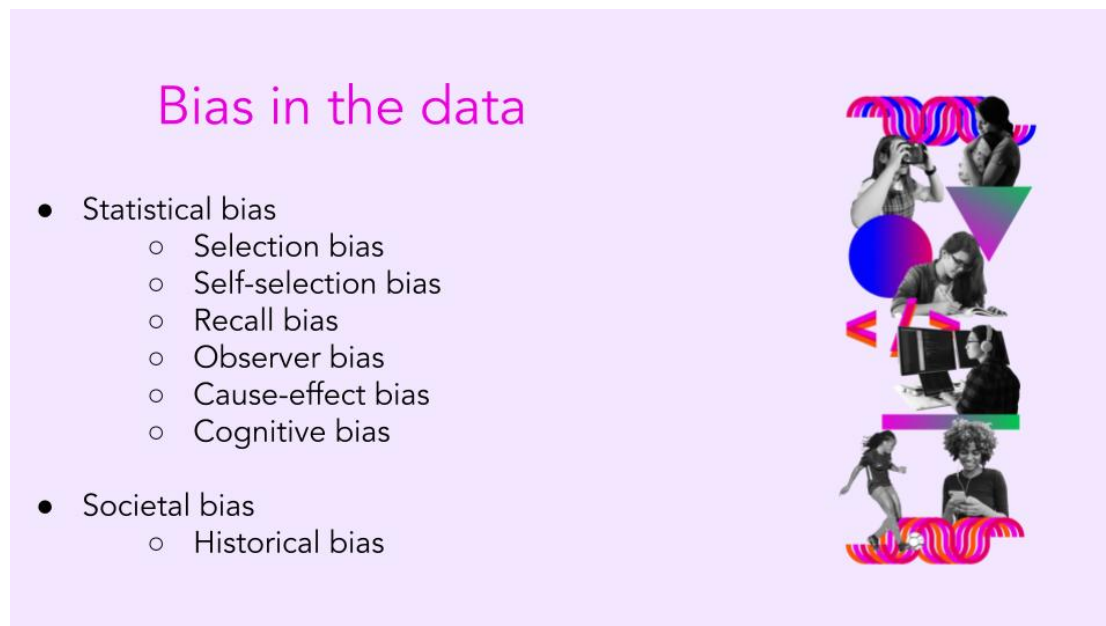


Figure 3.9 – Bias in the data can be separated into statistical bias (a systematic mismatch between the sample used to train a predictive model and the world as it currently is) and societal bias (inequalities of the society)

We ask about how the dataset was collected, over what timeframe, or if any information is sensitive. These questions are essential to understand the different types of biases. The appendix A.1 shows an example of different types of bias, the questions we ask to help identify them and proposed mitigation techniques.

A way to separate the bias in the data is by separating into bias into statistical (e.g. selection bias, observer bias) and societal (historical bias) (Figure 3.9). We use the definitions from the work *Prediction-Based Decisions and Fairness: A Catalogue of Choices, Assumptions, and Definitions* of authors (S. Mitchell et al., 2018). *Statistical bias* is a “systematic mismatch between the sample used to train a predictive model and the world as it currently is” (S. Mitchell et al., 2018). *Sampling bias* occurs “when a data set is not representative of the entire population to which the resulting model will be applied”. The nonstatistical notion of *societal bias* comes even if the training data are representative and accurate, as “they may still record objectionable aspects of the world that run counter to the decision-maker’s goals if encoded in a policy” (S. Mitchell et al., 2018).

For example, to identify measurement bias (the way that the training data is measured and collected can introduce inaccuracies in the model) (Suresh & Gutttag, 2020), we use the question from (Geburu et al., 2018)

- “What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?”

Mitigation: Pre-processing techniques

Pre-processing approaches tend to reconstruct the sample distributions of protected variables. In general, these approaches make specific alterations to the data to eliminate discrimination from the training data (S. Mitchell et al., 2018).

The goal is to “repair” the dataset and train the model on the new dataset.

This approach is considered the most flexible in the data science pipeline because it does not make assumptions on the choice of the machine learning algorithm that will be used.

Overall, addressing issues of societal bias is exceptionally challenging. It might require additional data collection or manually changing the model building process by understanding this bias. Or sometimes, this type of bias might not have a technical solution at all (S. Mitchell et al., 2018).

Between the many different approaches at the pre-processing level, we use the *Reweighting algorithm* (Kamiran & Calders, 2011) implemented with the IMB aif360 python package (Bellamy et al., 2018a).

The students discover the effectiveness of this method as the fairness metrics improve compared to the baseline model (Figure 3.10). Participants see the trade-off between group fairness (statistical parity and equal opportunity) vs individual fairness (generalised entropy) where the group metrics improved while the individual metrics got worse compared to the baseline model.

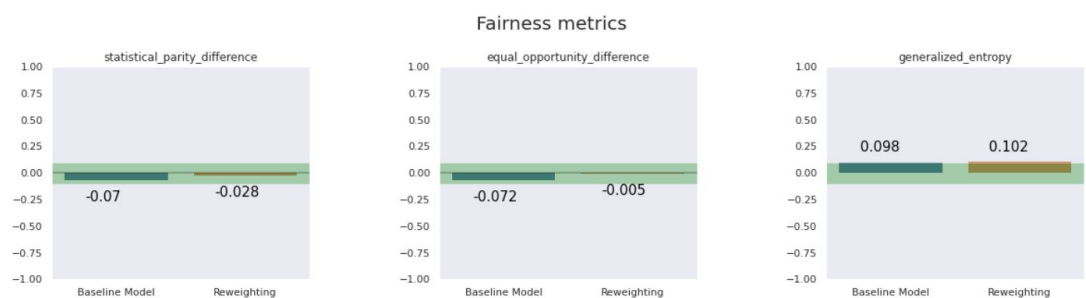


Figure 3.10 – Fairness metrics between the Baseline model and the fixed model with the Reweighting technique. Bias decreases when we compare group fairness metrics (statistical parity and equal opportunity), while it increases for the individual fairness metric (generalized entropy).

A discussion follows this session to compare the model’s results using the pre-processing algorithm with the baseline model.

3.4.7 In-processing

For this part, we present how bias is introduced in the design decisions made when creating the algorithm. Choices for building a model include the class of the model, the function and model parameters, among others (S. Mitchell et al., 2018).

Objectives

For this sections, participants are expected to:

- identify algorithmic bias
- try in-processing techniques to mitigate algorithmic bias

Content

The in-processing is about assessing the learning algorithm. Fair machine learning algorithms do not optimise for the fairest solution; they see fairness as a constraint on the set of feasible solutions.

Bias in Model

Bias does not derive only from training data. It can appear from (S. Mitchell et al., 2018):

- inappropriate data handling
- inappropriate model selection
- incorrect algorithmic design or application

To be clear about these choices, we use the questions from (M. Mitchell et al., 2018):

- “What type of model is it? This includes basic model architecture details, such as whether it is a Naive Bayes classifier, Convolutional Neural Network.”
- “Have we ensured that the model does not rely on variables or proxies for unfairly discriminatory variables?”

Although seemingly technical, these questions are related to the human rights principles of participation and accountability. Each decision (ex., the covariates we use to make a prediction) can introduce feature bias, meaning that a person could be excluded from our algorithm because of our choices.

Mitigation: In-processing techniques

The in-processing techniques focus on modifying the learning algorithm. They usually integrate a fairness penalty in the loss function, take the fairness metric as one of the inputs, and return a new classifier.

We need to explain and justify why a prediction or decision was made for an ethical decision-making process. It is currently not feasible with black-box models. Numerous in-processing approaches acknowledge that modelling techniques frequently become biased for various reasons: either by the dominant features, other distributional effects, or because they try to balance multiple model objectives, such as having an accurate and fair model. These approaches combine one or more fairness metrics into the model optimisation functions. The function converges towards a model parametrisation that maximises performance and fairness.

In the workshop, we use as an in-processing technique the *Meta Classification algorithm* (Celis et al., 2020). The meta-algorithm here uses the fairness metric as input and returns a classifier optimised for that fairness metric.



Figure 3.11 – Fairness metrics between the Baseline model and the fixed model with the Reweighting technique. Bias decreases when we compare group fairness metrics (statistical parity and equal opportunity), while it increases for the individual fairness metric (generalized entropy).

3.4.8 Post-processing

In the last part, we present the main assumptions made when we use the evaluations of the algorithmic output.

Objectives

For this section, participants are expected to:

- become familiar with the main assumptions that are made when we make decisions based on the algorithmic output
- try post-processing techniques to improve fairness

Content

Bias when making decisions

According to paper (S. Mitchell et al., 2018), when we are making the evaluations, there are three main assumptions :

1. Decisions are evaluated as an aggregation of independently evaluated decisions. This assumes that the decisions of others do not influence outcomes, an assumption known as no interference.
2. All individuals are considered symmetrically. This assumes that the harm of refusing someone, who could repay, a loan is the same among all people.
3. Decisions are evaluated at the same time in a batch instead of serially. That way, they miss possibly significant temporal dynamics.

In the workshop, we are using the loan application scenario. These assumptions are translated to the following examples to help participants see the effects. The first assumption implies that refusing one family member a loan could impact directly their ability to repay their loan. The second assumption means that rejecting a loan for education will perhaps affect that person's life in a very different way than it would affect a rejection of a similarly sized loan for a holiday house (S. Mitchell et al., 2018).

Participants see that decisions are not binary, but they have a serious implication on people's lives.

Mitigation: Post-processing techniques

Post-processors train on predictions from a black-box estimator and ground-truth values to produce fairer predictions.

In the workshop we use the *Calibrated Equalized Odds post-processor* (Pleiss et al., 2017). It combines the training and prediction of an arbitrary estimator and the post-processor while seamlessly splitting the dataset.

Again, as we see from Figure 3.12, the participants see the trade-off between group and individual fairness metrics. They see for themselves that pre-processing (Figure 3.10) and post-processing techniques (Figure 3.12) are the most flexible and practical approaches as they do not need to access the algorithms and machine learning models.

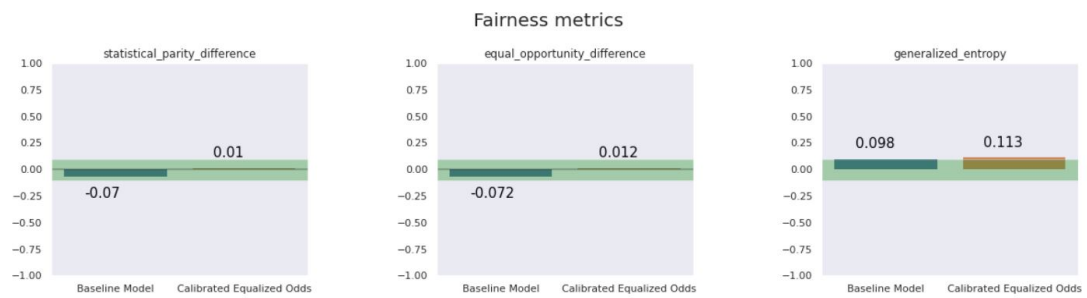


Figure 3.12 – Fairness metrics between the Baseline model and the fixed model with the Reweighting technique. Bias decreases when we compare group fairness metrics (statistical parity and equal opportunity), while it increases for the individual fairness metric (generalized entropy).

4 Evaluation of the framework

We evaluated our methodology in an iterative process of three workshops at two different Universities, École Polytechnique Fédérale de Lausanne (EPFL) and University College Dublin (UCD).

A workshop in 2020 (workshop #1) at EPFL, held just before COVID lockdown, was a success, particularly in moving the needle on understanding human rights frameworks, but lacked the technical components that students felt they needed in order to put the concepts into working reality.

This work is the rework in 2021 at EPFL (workshop #2). We again followed a qualitative approach to evaluate our methodology, combining questionnaires and extended in-depth user interviews. We used the findings of the first and second workshops to improve and enhance the workshop methodology.

We used the second workshop to verify the new, more technical framework and get feedback on our material, including suggestions for improvements and revisions.

During the third workshop, this time in a new university setting, UCD (workshop #3), we performed the revised workshop to validate the framework. We collected the results with questionnaires and semi-structured interviews with the members.

Workshop #1 2020 in real life (IRL) version fulfilled its maximum enrollment of 35 students (although 50 registered) for a full morning and lunch. (Table 4.1). Workshop #2 held virtually over Zoom in 2021, had 35 registrations, yet in the end, due perhaps to Zoom fatigue, we had 12 student participants mostly from technical backgrounds (75% doing a Masters' in Data Science and 25% doing a PhD in Digital Humanities).

Workshop	Place	Participants
#1	EPFL - 2020 (IRL)	50
#2	EPFL - 2021 (Zoom)	35
#3	UCD - 2021 (Zoom)	15

Table 4.1 – Places where we presented the workshop *<AI & Equality> A Human Rights Toolbox* and registered participants

4.1 Testing the methodology: Workshop #2 - EPFL

To attract students that were interested to learn more about this topic, we promoted the event widely at EPFL. We posted through the university portal, student mailing lists and courses.

The event took place on Zoom on Friday, 26th March 2021, in the afternoon, and it lasted for 3 hours.



Figure 4.1 – Poster for the workshop #2 at EPFL on Friday 26th March 2021

4.2 Participants

Workshop #2, held virtually over Zoom, had 35 registrations (Table 4.1), yet in the end, due perhaps to Zoom fatigue, we had 12 student participants.

From Figure 4.2, we see that most of the participants have a technical major. 66.6% belong in the more “hard science” fields (Data Science, Computer Science, Electrical Engineering), while the rest (33.3%) is from Digital Humanities. Also, the majority of participants is from the master level since the participants needed to have more of a machine learning background (which at EPFL it is taught at the master level). In terms of gender, we had seven participants

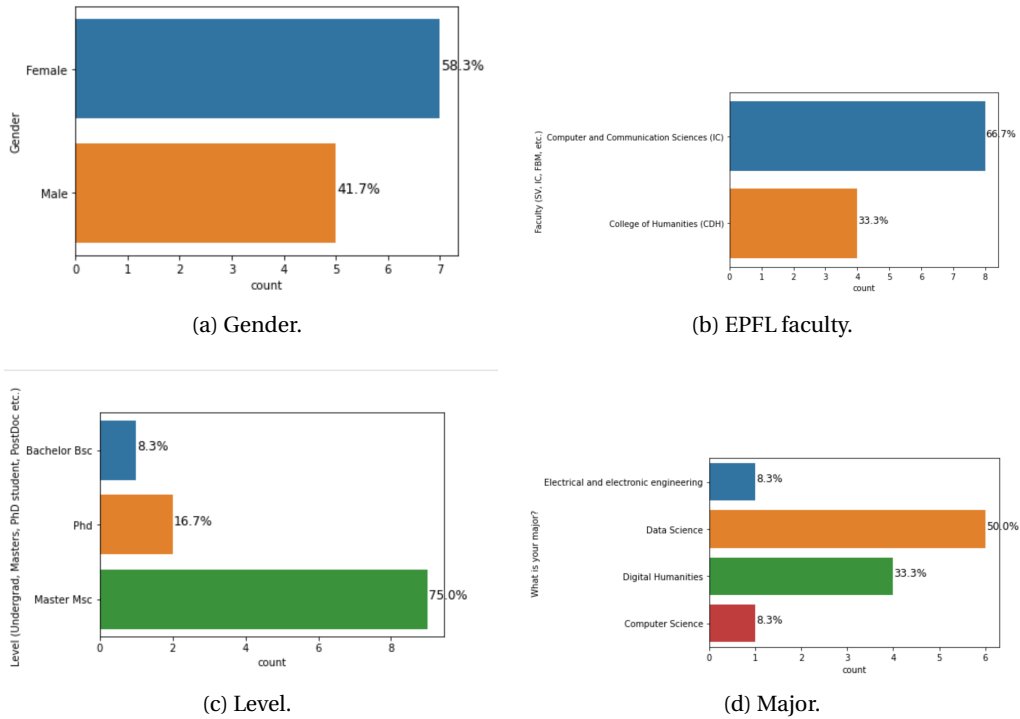


Figure 4.2 – Demographic plots and student status of the EPFL workshop participants

who identified as female and five as male.

After the workshop, we did a follow-up semi-structured interview with a participant from Data Science, as well as a structured debriefing with our colleague from Office of the United Nations High Commissioner for Human Rights (OHCHR).

4.3 Questionnaire

We used both close-ended and open-ended questions to evaluate the workshop. The complete questionnaire of the workshop is available in the annex A.2.

To collect data, we used Likert-type (Likert, 1932) statements and a continuum of possible responses, with a 5-point semantic differential rating scale (Ranging from Strongly-disagree to Strongly Agree). Each item is given a numerical score so that the data can be analyzed quantitatively. It is inspired by PSSUQ (Post-Study System Usability Questionnaire) (Lewis, 1992). The participants were asked to evaluate the length of the workshop, the usefulness, the information quality. To calculate the overall satisfaction of the users from a score out of 5, we used the following formula (where i is the question number):

$$\frac{\sum_{i=1}^{i=\# \text{ of questions}} score_i}{\# \text{ participants} \times \# \text{ of questions}}$$

4.4 Feedback

The main categories of the comments we received were about the Practicum, the time schedule, and the event's interactivity.

The comments we received from the participants focused on the content and the Practicum. They found it challenging to concentrate either on the rich text of the Jupyter notebook or on the speaker. Regarding the Jupyter notebook, participants mentioned that “The notebook is a very good idea as it helps with engaging in the workshop and contains all the information we might want to revisit later on. However, I sometimes felt a bit confused about whether I had to follow what the presenter was saying or focus on reading the script. I feel like the presentation part should be supported by clear slides containing the bare minimum, and then we can revisit the content in a more detailed way in the notebook if we wish.”. Another participant proposed that it “Could be useful to ask the participants to explore the notebook before, and then present the notebook during the session and explain it”.

Finally, a participant with a computer science background mentioned: “The ‘debiasing’ techniques were not covered in great detail, so apart from seeing the results in the notebook, I could not say which is preferable, how they work, and what their limitations are. But it’s a technical issue, so I guess I didn’t expect to be an expert by the end.”

The comments from the other organizers and colleagues were mainly about the time scheduling and the interactivity in an online context. The many small breaks between the practical session resulted in students losing focus and made the event last longer than it needed.

The colleague from OHCHR proposed: “To enable more interactive methods, we could consider asking each participant to share one takeaway message from the workshop. If you don’t interact actively, you don’t retain what you gained.”

From the interview with the data science participants, we got a more in-depth comment of what could be improved in the Practicum. The participant wanted to see “More links between the human rights part and the algorithms”, and “stronger, explicit links and explanation how and why a practice violates human rights”.

On the positive parts of the workshop was the “different fairness metrics and how they are not compatible with each other” (impossibility theorem), as well as the chance to see the different sources and types of bias, and how they “enter” the system.

However, she expected to write more code and to focus more on the interpretation. She proposed to “Walk through the first example and explain the method and the code”, and focus on “explaining decisions: why we decide this, what are the implications, how they change the results”. She also agreed that giving the notebook will help participants to “get familiar”.

4.5 Specifications

From the comments collected from the participants and the colleagues, the main issues focused on the amount of information that we presented at the workshop.

Below is a comprehensive list of issues that could be used as improvements for the following workshop versions. The issues were focused on the content and the timing.

Content

1. Create slides to support the teaching material
2. Focus on the explanations: go deeper in explaining the decisions, the implications, the motives
3. Create a different pattern for different parts of the presentation: change between explanations, case studies and duration of each session
4. Create a case study, throughout the workshop

Timing

5. Give the Jupyter notebook in advance so that participants become more familiar with the terms
6. Reduce the amount and the duration of the breaks

For the next workshop iteration, we decided to implement the improvements that would bring the most value to the participants. We addressed two issues regarding the content (specifications 1 and 2) and the two issues regarding the timing (specifications 5 and 6).

5 Validation of the framework

5.1 Validating the methodology: Workshop #3 - UCD

For the validation of the framework, the third iteration of the workshop was performed at University College Dublin (UCD), on May 20th 2021.

The format had some changes, compared to the Workshop #2 at École Polytechnique Fédérale de Lausanne (EPFL). First, the duration was decreased from 3 hours to 2 because of time constraints from the hosting university. This led us to remove the *Applied Research* part, where a representative (PhD, post-doc or professor) from the hosting university presents their work linking human rights and algorithms (methodology described in section 3.3 Applied Research). Also, the amount and the duration of the breaks decreased.

In addition, we applied the specifications that emerged from Workshop #3. Based on the feedback from the participants, we presented the theory in the practical toolbox of the workshop using slides, gave the notebook to the participants two days beforehand, and focused more on the explanations.

5.2 Participants

At UCD we recruited the participants through contact made directly by the hosting professor to her recent students. We had a total of 12 participants, both from technical and social science backgrounds (Figure 5.1). As with the EPFL participants, seven identified as female and five as male. Their education level was split almost half (6) and half (5) between masters students and PhD candidates. There was also one academic faculty member. Most of the participants had a social science education (digital humanities, digital policy, digital journalism, communication and media), while only three were from a technical background (computer science, computational linguists).

This difference in the participants from EPFL and UCD was indicative of the range that our workshop applies to. We believe that their background is a promising model for the future and

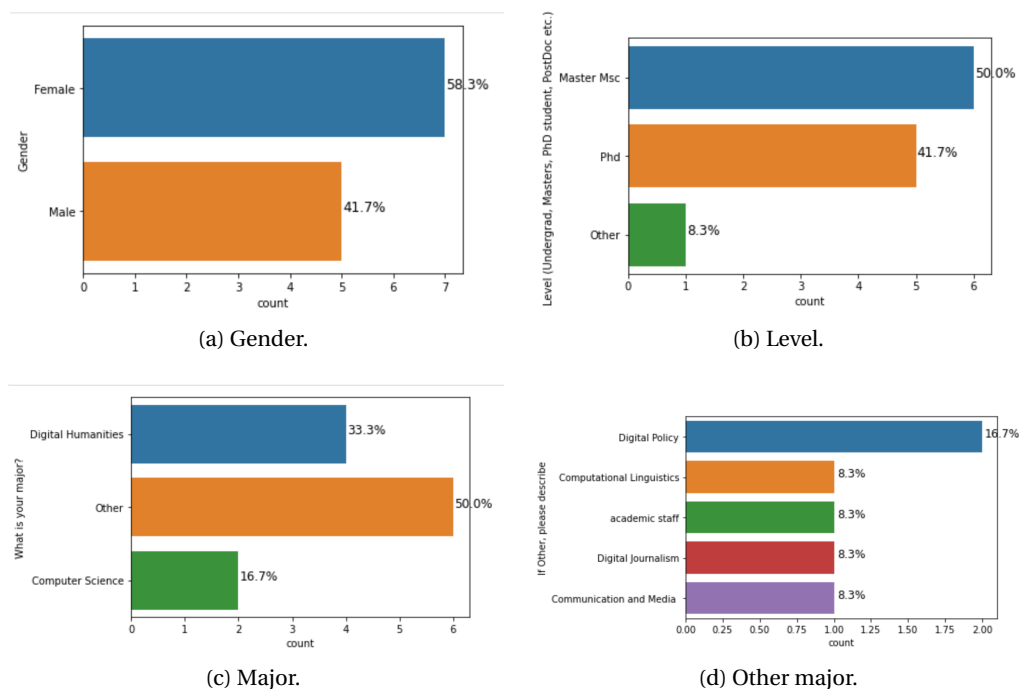


Figure 5.1 – Demographic plots and student status of the UCD workshop participants

the future of multi-disciplinary collaborations.

5.3 Results of Questionnaire

Participants responded very positively to the workshop, as could be seen from the results of the questionnaire (Annex A.2). The overall satisfaction from Workshop #2 increased from 89% to 4.7/5 = 98% score to the Workshop #3. The participants from Workshop #3 achieved the objectives of the workshop as seen from Table 5.1.

Objective	Before workshop (/5)	After workshop (/5)	Increase (%)
1. I would rate my confidence in describing the key elements of a human rights based approach to AI	2.5	4.5	80
2. I would rate my confidence in describing the equality and gender, racial and other forms of discrimination relevant to the design of algorithms	3.125	4.5	44
3. Overall, I would rate my ability to identify the relevance of different biases and the importance of gender, race and equality to computer science and engineering	3.75	4.75	26
4. I would rate my ability to analyse how gender, racial and other bias has occurred or can occur in the research, design and development of AI	3.25	4.25	30
5. I would rate my ability to use tools and techniques to mitigate bias in AI	2.375	3.625	52
6. I would rate my ability to evaluate methods to integrate non-discrimination into design, planning and implementation of AI projects	2.5	4	60

Table 5.1 – The evaluation of the objectives by the participants from Workshop #3 at UCD

Besides the scores of the objectives that were reached (Table 5.1), many participants claimed to understand the connection between Human Rights and AI and the complexity of fairness. On the key take-aways, a participant wrote: “Human rights should be embedded into all aspects of AI design, not an ‘add-on’”, while another said, “Fairness is a complex concept and is context-sensitive.”

The value of interdisciplinarity between sectors was also recognized at the workshop. One participant explained, “I really like the interdisciplinary perspective and the encouragement that people were given to ask questions and present their own take on the learning content.”. Another participant agreed, mentioning that “It was great to have the opportunity to hear from everyone and to speak during the session. I personally felt more comfortable contributing, knowing that we were encouraged to do so, and hearing about how other people will apply the content to their work made the session even more enjoyable.”

Regarding the content, although we tried to keep it at a high level, “The material felt approachable, and the practical usefulness of the content was clear throughout, so it never felt overly-theoretical”, as one participant said.

Going forward, to apply the skills they learned from the workshop, time and practice were mainly mentioned. A participant said “I would need more time to go deeper into the notebook and manipulate data/code myself to gain a deeper understanding of the impact that the pre-processing decisions/different algorithms have on data and outcome.”

5.4 Analysis of results

From the feedback of the participants, we analyzed how the participants reacted to the new specifications.

1. Create slides to support the teaching material

Having slides was a significant aid for this faster-paced presentation of the workshop. Participants reacted very positively to it, as it could be seen from their comments “The slides were impeccable and the notebook is very well done and easy to understand.”. Also, since a participant said that “As a person from a non-technical background, I found the presentation very interesting and enlightening.”

An improvement would be to present the time allocation when starting the workshop. A participant commented: “At the very beginning, it would have been nice to have a roadmap of who was going to present what and in what order, how much time would be allocated to the different sections, whether the slides would be available or not. ”.

2. Give the Jupyter notebook in advance so that participants become more familiar with the terms

We emailed the Jupyter notebook to the participants 2 days before the workshop. That change had an immediate effect as one participant reported: “I found it really helpful to have access to the notebook in advance. This gave me a chance to become a little bit more familiar with the terms prior to the presentation, and I think it helped me to follow along.” They continued “I think the email mentioned that it wasn’t mandatory to look at the notebook in advance, but I think that for the 2-hour session where we move through the content more quickly, this could be changed to being recommended.”.

This addition was beneficial, and it will be kept for future versions of the workshop.

3. Focus on the explanations: add more text and more visuals to aid with the explainability of the decisions

For this part, we added more comments based on the feedback from the participant’s interview in Workshop #2. When explaining the different metrics, we added why the confusion matrix is essential, how fairness measures derive from the confusion matrix and how pre-processing decisions affect different metrics. We also expanded the exploratory data analysis with more plots and analyzed the performance metrics in greater detail.

A participant mentioned “The session could possibly allow for a little bit more time to cover the confusion matrix section. I understand that this was a shorter version of this presentation than usual, though, so, within the time that we had, I think it was all covered very well, and it was extremely helpful - thank you”. Another participant also agreed, saying that “My only suggestion is that it might be helpful to allocate more time to the data processing section, but I understand that you were working within a shorter timeframe than usual. ”

From the comments, we conclude that, although the explanations were improved from the previous workshop, we still need more time to make stronger connections between the different metrics and going into more depth about how each decision affects the outcome. A proposal would be to allocate enough time on the first pre-processing part, following an even more detailed step-by-step approach of how a “good practice” data analysis could have been.

4. Reduce the breaks

Although reducing the time for the breaks helped more concentrated participants, the general reduction from 3 to 2 hours affected the information flow. A participant commented “At this session, a 3-hour workshop was squeezed into 2 hours, which became more clear as we neared the end of the workshop. The organizers could, therefore, either keep the original 3 hours (which would be beneficial in my opinion) or cut bits and pieces (e.g. by eliminating question sessions after each part) to make it fit into 2 hours.”.

A participant also agreed to say that “I think there might have been 30 minutes for more questions ”.

Changing the time duration was noticeable. We see that participants needed the extra time, and proper time allocation should be considered for a future workshop.

Conclusion

In general, participants were very positive about the workshop. Some of them shared their excitement in the comments, mentioning that “I’ll make sure to follow Sofia’s project and to keep this toolbox in mind for my future research.” The material proved helpful to other participants, reporting that “I think for the purpose of completing my MSc thesis, the materials provided here today and the extra resources listed in the notebook will have me well covered.”, and “Thank you again for your workshop today, it was extremely helpful, informative and enjoyable”.

Finally, going further, a participant proposed that “It could be expanded to look at its impact on the public sector”.

Overall, besides the comments and proposals for future improvements, participants responded well to the workshop, validating our methodology and proving that our work is essential and timely.

6 Discussion

In line with the research questions, in this thesis, we explored

“How can International Human Rights frameworks be integrated with current concepts of fairness for the design of an educational tool for computer scientists?”

Through a series of workshops and semi-structured interviews, we verified that our methodology *<AI & Equality> A Human Rights Toolbox* gave the participants the skills to describe the critical elements of a human rights-based approach to AI, identify and analyse different biases in research, and use tools and techniques to mitigate bias in AI.

In this section, we summarise our results and discuss our contribution, the limitations of our current methodology and future directions for the project.

6.1 Summary

This project presents the methodology that includes a workshop consisting of a Human Rights module and code, outreach, and community plan that incorporates human rights concepts with data science and integrates international human rights frameworks with current concepts of fairness. The goal is to create a transformational educational tool for computer science students.

Research Objective 1: Study current approaches on ethics and fairness in machine learning and examine their main limitations

Regarding the first research objective, we analysed how ethics and fairness is applied in machine learning and concluded that most approaches focus on applying a mathematical definition of fairness. However, ethics and fairness are socio-cultural concepts, and a solely technical solution is an oversimplification of the complexity of real-world problems.

We address this limitation in the tool's design using a human rights-based framework, as human rights are often better defined and measurable, and most of them are defined under international or national law.

Then we analyse the AI ethics principles published by many different organisations to help guide a more ethical development of machine learning systems. However, these approaches are still very high level, and it is difficult for partitioners to use them in their daily tasks. On the other hand, we also examined several fairness toolkits that transform these principles into code. Their limitation comes from the complexity of the task they are trying to solve. It is a complex problem: fairness does not have a single definition, and much reading is needed to understand what should be applied. Just as the mathematical definitions, these technical approaches fail on more challenging real-life problems.

We identify the difficulty between transforming the theory into practice. We address this issue by doing the transformation step by step. Starting with the human rights principles, we see the questions we need to ask ourselves before writing code, and then we show how we can analyse and create an algorithm in a human rights-based approach.

We also examine how certain universities approach the teaching of tech ethics and whether it is included in their curriculum. We see that most successful approaches (Grosz et al., 2019) (Choirat et al., 2020) are the ones that are based on interdisciplinarity and merge social with technical sciences. We identify that one of the problems comes from the approach of different universities: having ethics as an on-off course and not as part of student's work, students fail to the importance of it.

To overpass this, we created it in a workshop to be easier for a student; a few hours are enough to cover the material. We used the Jupyter notebook to host the workshop's material and present the analysis, as it is the most used tool in the exercise sessions for data scientists.

Research Objective 2: Propose, design and build an educational tool for computer scientists

Regarding the second objective, the evaluation and the validation showed that the educational tool achieved its purpose.

The limitations of the previous approaches and the gaps we identified helped build a workshop that uses a human rights-based framework when designing and building a machine learning algorithm.

We incorporated a human rights-based approach in the fairness pipeline: data, model and evaluation by separating each part into sections: where bias can be found, what questions practitioners should ask that related and finally, mitigation techniques for the bias.

The workshop has already been conducted at University College Dublin (UCD) and École

Polytechnique Fédérale de Lausanne (EPFL) (through an iterative process to improve the offering each time in response to participant feedback). These workshops showed that students improved their understanding of human rights concepts and their ability to identify and analyse how gender, racial and other bias can occur in the research design and development of AI and identify and use tools and techniques to mitigate bias in AI.

6.2 Contribution

Our work identifies and tries to solve a significant problem in computer science from a unique lens – that computer scientists are often unaware of basic human rights concepts even as they engage in issues having to do with human rights.

Our novelty comes from the development and collaboration with human rights experts from the Office of the United Nations High Commissioner for Human Rights (OHCHR). This is the first and to date only foray of OHCHR into the academic world of computer science students to jumpstart a conversation about a human rights-based approach being the baseline from which we should create new algorithms and new models. It provides actionable steps for universities to integrate a human rights framework into their instruction of the next generation of computer scientists. The project has already begun to gain traction, having already refined and validated the workshop offering at various universities and beginning to measure impact on participants.

The novelty comes from the conceptual perspective, combining experts from different disciplines, each contributing knowledge to the project. Using human rights as a starting point, we explicitly integrate an international human rights framework into computer and data science — as a means to mitigate persistent challenges surrounding fairness. The nature of the intervention of our methodology is inherently interdisciplinary. It is administered not only by computer science professors but also experts in law/philosophy/ethics and people in social sciences. This allows incorporating matters concerning the law.

The general effort of raising awareness on the societal implications of AI is relevant to the computer science community, the students, the universities and society. The goal is to create an inter-disciplinary event, so the problem does not get oversimplified as a technical one.

6.3 Limitations of the methodology

This work is the first step toward incorporating a human rights-based approach in a computer science course.

Evaluation and validation

The biggest challenge of this approach comes from the difficulty to validate such educational tools. It is hard to capture to what degree students learn the critical concepts that the workshop is trying to teach.

The concept of human rights, fairness and changing the way participants have been taught to approach machine learning algorithms needs time and probably repetition to sink into the students. It is a process that takes time to be adopted and internalised. This is not easy to evaluate, and it goes beyond the scope of this study.

Content

As seen from the validation results and the comments received, the material is rich yet quite heavy. Although we do not go into much depth, the time allotted has not been enough to cover all the questions asked. In future versions, we will consider splitting and spreading the workshop into different days to maximise the time for discussion and interaction.

Organisation

As this is an initial effort, the data is limited for quantitative analysis to have a statistically significant result.

The methodological choices were constrained by the time each iteration of the workshop needed. It requires human resources (availability from the presenters, outreach and contact with the hosting university and its researchers, outreach to and interest of students) and organisation.

6.4 Recommendations for future work

Our future work going forward is related to the three essential components of our methodology: help to nourish a community, blended workshops and self-guided material.

Community outreach

One of the most important parts is the creation of the community. We want to engage, support and nourish a genuinely international community of university students in computer science that want to make a difference. These students often do not have the tools or university structure to do AI for good or the infrastructure to reach out and collaborate with students from other disciplines who would want to make a difference in the algorithmic realm, such as those studying public policy, for example.

We need a genuinely interdisciplinary approach to solve this problem. Social scientists have theoretical knowledge, and they are aware of most types of social biases. They can identify pitfalls when dealing with social data, but they could lack the technical skills, especially when it comes to big data. We can provide the community, through which they can connect with data scientists to perform applied research. The Jupyter notebook is also an accessible format for non-technical students to understand how coding works, as it was demonstrated at Workshop #3 at University College Dublin (UCD).

On the other hand, data and computer science students have the technical knowledge, and most are the ones who perform the analysis. However, most of them are trained in a “hard science” background. They often have limited or insufficient domain knowledge, or they could even not grasp the social effects that data and algorithms have on people. By connecting with social scientists, they can design a project that is aware of the limitations and the effects on society.

In addition, by being a truly international community, students from one region can use and interact with students from other regions. They could use their local data, understanding the local socio-cultural context from other students in the first person.

With this community, we can create different working groups between the different profiles of the students. Law and social science students present why an algorithm in the “wild” can create problems, helping computer science students identify and understand their work’s effects. We need scientists and engineers that understand the intersectional dimensions of their work and the implications that work has for all citizens because, as scientists, we have a unique potential to make a social impact in the real world.

Blended workshops

We plan to expand and do the workshop in other universities. The blended workshops will be done in conjunction with different universities or different departments within a university engaging. They call the academic community to add more legal/ethical/social science material and help understand how code can potentially create insights and solutions. We want academics from a technical and law/human rights/social sciences background to perform the blended workshops jointly.

The universities’ top legal and human rights scholars with top data science machine learning professors can create content for their sessions. Under a creative commons license, they can add their insights and lectures to the AI Equality Toolbox.

Evaluation

Another way to go forward for the next version is to address the limitation of the evaluation and validation of the tool.

There are several ways to advance. One could be an additional in-depth interview with more participants to dig deeper into their opinions. We could collect additional comments from other academics from their participation in the blended versions of the workshop. A follow-up with the participants could show probably the impact the workshop had on their work.

Another approach is to design and perform an experiment to evaluate the tool as an intervention. A potential experiment design is to have two groups, one that does the workshop and one that does not. After the workshop, they could be evaluated on a data science task, either a use case evaluation on the discussion and specifications they give to the problem.

Expand to public sector

Finally, the approach for future work is expanding this type of workshop to the public sector. The idea was part of the vision of the Women at the Table (W@tt) for the *Human Rights-Based Approach*, but it was also proposed by a participant from the UCD workshop who studied digital policies.

We want to create a similar light-touch approach for policymakers, including terms and material commonly used in automated or algorithmic decision-making processes. The goal is for technologists and policymakers to find a shared vocabulary for the design and deployment of these conversations.

7 Conclusion

By performing a series of workshops, this thesis has shown how a human rights-based approach can be integrated with current concepts of fairness for designing an educational tool for computer scientists.

We wish to bring an international university generation to understand the scientist's unique potential of social impact in the real world, bridging science and human rights policy to foster systemic resilience and more equal, just, robust democracies.

With the *<AI & Equality> A Human Rights Toolbox* we found that students improved their knowledge and understood the importance of human rights in their code. Going forward, we want to create a space and content for young policymakers and young scientists to gather and find resources and one another to create the technology we need and the technology we deserve, in line with the human rights values we all embrace.

Closing with a quote from Michelle Bachelet, the UN High Commissioner for Human Rights, in her keynote speech for “Human rights in the digital age” (Bachelet, 2019)

“ To respect these (human) rights in our rapidly evolving world, we must ensure that the digital revolution is serving the people, and not the other way round. We must ensure that every machine-driven process or artificial intelligence system complies with cornerstone principles such as transparency, fairness, accountability, oversight and redress.”

This should start with the current generation of computer scientists.

A An appendix

A.1 Biases

A list of different types of biases that can occur when collecting data or designing and creating an algorithm.

Bias	Funding	Historical
Definition	results are reported in order to support or satisfy the funding agency or financial supporter of the research study	Already existing bias and socio-technical issues in the world and can seep into from the data generation process even given a perfect sampling and feature selection
Example	Testing methods, when paid for by for-profits, may have an inherent bias towards producing positive outcomes for the company.	Developing an algorithm that uses ZIP code as a feature in predicting hospital length of stay, for use in assigning a case manager to those who are predicting to stay a shorter amount of time. Results in discrimination against socioeconomically disadvantaged patients
Question	Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.	Does the dataset relate to people? Does the dataset identify any subpopulations (e.g. by age, gender)? Does the dataset contain data that might be considered sensitive in any way?
Mitigation	Researchers should retain control over the design, conduct, analysis, and reporting of the study, especially avoiding research contracts which include non-disclosure agreements or that allow the sponsor to have any role in the design, conduct or publication of the research. All data from the study should be publicly available.	Developers should work with ethicists and all stakeholders to consider the implications of using particular dataset features

Table A.1 – Example of different types of bias (Funding and Historic). Each bias is defined with an example and a mitigation technique. All questions come from *Datasheets for Datasets* (Gebru et al., 2018) and *Model Cards for Model Reporting* (M. Mitchell et al., 2018)

Bias	Sampling	Temporal
Definition	Occurs due to non-random sampling of subgroups. The trends estimated for one population may not generalize to data collected from a new population	From differences in populations and behaviors over time. Occurs when we assume a wrong sequence of events which misleads our reasoning about causality.
Example	A voice recognition technology is trained only with the audio language data generated by individuals with a British accent. This model will have difficulty in voice recognition when an individual with an Indian accent interacts with it. This results in lower accuracy in performance.	An example can be observed in Twitter where people talking about a particular topic start using a hashtag at some point to capture attention, then continue the discussion about the event without using the hashtag
Question	If the dataset is a sample from a larger set, what was the sampling strategy (e.g. deterministic, probabilistic with specific sampling probabilities)?	Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g. recent crawl of old news articles)? In not, please describe the timeframe in which the data associated with the instances was created. List when the dataset was first published
Mitigation	Using random methods when selecting subgroups from populations. Ensuring that the subgroups selected are equivalent to the population at large in terms of their key characteristics	-Ask for information about the temporal sequence in your questionnaire. - Use a prospective study design

Table A.2 – Example of different types of bias (Sampling and Temporal). Each bias is defined with an example and a mitigation technique. All questions come from *Datasheets for Datasets* (Gebru et al., 2018) and *Model Cards for Model Reporting* (M. Mitchell et al., 2018)

Bias	Representation
Definition	Occurs from the way we define and sample from a population. The training set does not accurately represent real-world data
Example	Training a melanoma classifier to detect cancer for patients with white skin only, and then expecting it to perform well on darker skin colors
Question	Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g. geographic coverage)? Is so, please describe how this representativeness was validated/verified. When generating a more representative cohort would either entail downsampling the existing records to achieve population representation or to spend more money to actively seek out patients that have slipped through the cracks
Mitigation	<ul style="list-style-type: none">- Establish best practices in dataset collection- Encourage diverse dataset development

Table A.3 – Example of different types of bias (Representation). Each bias is defined with an example and a mitigation technique. All questions come from *Datasheets for Datasets* (Gebru et al., 2018) and *Model Cards for Model Reporting* (M. Mitchell et al., 2018)

A.2 Questionnaires

Question	Available Answers					
Gender	M	F	X			
Major	Computer Science	Data Science	Digital Humanities	Other (define)		
Level	Bachelor Bsc	Master Msc	Phd	Postdoc	Other (define)	
The presentations were the right length	1 (Strongly Disagree)	2	3	4	5 (Strongly Agree)	
There was enough time to ask questions and seek further information	1 (Strongly Disagree)	2	3	4	5 (Strongly Agree)	
Presenter A (OHCHR representative) was knowledgeable about the subject matter	1 (Strongly Disagree)	2	3	4	5 (Strongly Agree)	
Presenter A (OHCHR representative) was able to effectively communicate their information	1 (Strongly Disagree)	2	3	4	5 (Strongly Agree)	
Presenter B (Data Science representative) was knowledgeable about the subject matter	1 (Strongly Disagree)	2	3	4	5 (Strongly Agree)	
Presenter B (Data Science representative) was able to effectively communicate their information	1 (Strongly Disagree)	2	3	4	5 (Strongly Agree)	
All workshop materials supported the presentation well	1 (Strongly Disagree)	2	3	4	5 (Strongly Agree)	
The notebook was easy to follow	1 (Strongly Disagree)	2	3	4	5 (Strongly Agree)	
Further comments about the delivery of the notebook, presentations and length of sessions	open question					
I would rate my confidence in describing the key elements of a human rights based approach to AI (before event)	1 (not aware)	2	3	4	5	Don't know
I would rate my confidence in describing the key elements of a human rights based approach to AI (after event)	1 (not aware)	2	3	4	5	Don't know
I would rate my confidence in describing the equality and gender, racial and other forms of discrimination relevant to the design of algorithms (before event)	1 (not aware)	2	3	4	5	Don't know
I would rate my confidence in describing the equality and gender, racial and other forms of discrimination relevant to the design of algorithms (after event)	1 (not aware)	2	3	4	5	Don't know
Overall, I would rate my ability to identify the relevance of different biases and the importance of gender, race and equality to computer science and engineering (before event)	1 (not aware)	2	3	4	5	Don't know
Overall, I would rate my ability to identify the relevance of different biases and the importance of gender, race and equality to computer science and engineering (after event)	1 (not aware)	2	3	4	5	Don't know
I would rate my ability to analyse how gender, racial and other bias has occurred or can occur in the research, design and development of AI (before event)	1 (not aware)	2	3	4	5	Don't know
I would rate my ability to analyse how gender, racial and other bias has occurred or can occur in the research, design and development of AI (after event)	1 (not aware)	2	3	4	5	Don't know
I would rate my ability to use tools and techniques to mitigate bias in AI (before event)	1 (not aware)	2	3	4	5	Don't know
I would rate my ability to use tools and techniques to mitigate bias in AI (after event)	1 (not aware)	2	3	4	5	Don't know
I would rate my ability to evaluate methods to integrate non-discrimination into design, planning and implementation of AI projects (before event)	1 (not aware)	2	3	4	5	Don't know
I would rate my ability to analyse how gender, racial and other bias has occurred or can occur in the research, design and development of AI (after event)	1 (not aware)	2	3	4	5	Don't know
Further comments about your learning and skills	1 (Strongly Disagree)	2	3	4	5 (Strongly Agree)	
I was satisfied that I was able to actively engage in the workshop	1 (Strongly Disagree)	2	3	4	5 (Strongly Agree)	
I was satisfied that my views and information I shared have been heard	1 (Strongly Disagree)	2	3	4	5 (Strongly Agree)	
Final comments about participation	1 (Strongly Disagree)	2	3	4	5 (Strongly Agree)	
I was satisfied with the workshop	1 (Strongly Disagree)	2	3	4	5 (Strongly Agree)	
I benefited from attending the workshop	1 (Strongly Disagree)	2	3	4	5 (Strongly Agree)	
The discussions were useful	1 (Strongly Disagree)	2	3	4	5 (Strongly Agree)	
I would recommend this workshop to other students	1 (Strongly Disagree)	2	3	4	5 (Strongly Agree)	
How relevant and helpful do you think it was for your work?	open question					
What were your key take aways from this event?	open question					
What parts of the workshop aided your learning the most?	open question					
Any comments and suggestions for improving this workshop	open question					
What did you like about the workshop?	open question					
What did you *not* like about the workshop?	open question					
What other tools, aid, support or opportunities would you feel you need to apply the skills you learned today in your work?	open question					

Table A.4 – The survey used for the evaluation of the workshop

Bibliography

- Adebayo, J. A. (2016). *Fairml: toolbox for diagnosing bias in predictive modeling signature redacted*. Massachusetts Institute of Technology. <https://dspace.mit.edu/handle/1721.1/108212>
- Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., & Wallach, H. (2018). A reductions approach to fair classification.
- Ajunwa, I., & Greene, D. (2019). Chapter 3 platforms at work: automated hiring platforms and other new intermediaries in the organization of work. <https://doi.org/10.1108/S0277-283320190000033005>
- Anderson, R. E. (1992). Acm code of ethics and professional conduct. *Communications of the ACM*, 35, 94–99. <https://doi.org/10.1145/129875.129885>
- Angwin, J., Larson, J., Kirchner, L., & Mattu, S. (2016). Compas analysis. <https://github.com/publica/compas-analysis>
- Arnold, M., Bellamy, R. K. E., Hind, M., Houde, S., Mehta, S., Mojsilovic, A., Nair, R., Ramamurthy, K. N., Reimer, D., Olteanu, A., Piorkowski, D., Tsay, J., & Varshney, K. R. (2019). Factsheets: increasing trust in ai services through supplier's declarations of conformity.
- Arya, V., Bellamy, R. K. E., Chen, P.-Y., Dhurandhar, A., Hind, M., Hoffman, S. C., Houde, S., Liao, Q. V., Luss, R., Mojsilović, A., Mourad, S., Pedemonte, P., Raghavendra, R., Richards, J., Sattigeri, P., Shanmugam, K., Singh, M., Varshney, K. R., Wei, D., & Zhang, Y. (2020). Ai explainability 360: hands-on tutorial. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 696. <https://doi.org/10.1145/3351095.3375667>
- Asako Hattori. (2020). A human rights-based approach: example. https://aiequalitytoolbox.com/assets/pdf/Introduction_to_Human_Rights.pdf
- Bachelet, M. (2019). *Human rights in the digital age - can they make a difference?* <https://www.ohchr.org/EN/NewsEvents/Pages/DisplayNews.aspx?NewsID=25158&LangID=E>
- Barocas, S., Hardt, M., & Narayanan, A. (2019). *Fairness and machine learning* [<http://www.fairmlbook.org>]. fairmlbook.org.
- Becker, B. (1994). UCI machine learning repository. <https://archive.ics.uci.edu/ml/datasets/adult>
- Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K. N., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K. R., & Zhang, Y. (2018a). *Ai fairness 360: an extensible toolkit*

- for detecting, understanding, and mitigating unwanted algorithmic bias. <https://github.com/ibm/aif360>
- Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K. N., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K. R., & Zhang, Y. (2018b). AI Fairness 360: an extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. <https://arxiv.org/abs/1810.01943>
- Bender, E. M., & Friedman, B. (2018). Data statements for natural language processing: toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6, 587–604. https://doi.org/10.1162/tacl_a_00041
- Benjamin, M., Gagnon, P., Rostamzadeh, N., Pal, C., Bengio, Y., & Shee, A. (2019). Towards standardization of data licenses: the montreal data license.
- Berk, R., Heidari, H., Jabbari, S., Joseph, M., Kearns, M., Morgenstern, J., Neel, S., & Roth, A. (2017). A convex framework for fair regression.
- Berk, R., Heidari, H., Jabbari, S., Kearns, M., & Roth, A. (2017). Fairness in criminal justice risk assessments: the state of the art.
- Bietti, E. (2019). *From ethics washing to ethics bashing: a view on tech ethics from within moral philosophy*. <https://papers.ssrn.com/abstract=3513182>
- Binns, R. (2021). Fairness in machine learning: lessons from political philosophy.
- Bird, S., Dudík, M., Edgar, R., Horn, B., Lutz, R., Milan, V., Sameki, M., Wallach, H., & Walker, K. (2020). *Fairlearn: a toolkit for assessing and improving fairness in ai* (tech. rep. MSR-TR-2020-32). Microsoft. <https://www.microsoft.com/en-us/research/publication/fairlearn-a-toolkit-for-assessing-and-improving-fairness-in-ai/>
- Birnbacher, D. (1999). Ethics and social science: which kind of co-operation? *Ethical Theory and Moral Practice*, 2(4), 319–336. <http://www.jstor.org/stable/27504102>
- Bonde, S., & Firenze, P. (2013). *A framework for making ethical decisions | science and technology studies*. <https://www.brown.edu/academics/science-and-technology-studies/framework-making-ethical-decisions>
- Buolamwini, J., & Gebru, T. (2018). Gender shades: intersectional accuracy disparities in commercial gender classification. In S. A. Friedler & C. Wilson (Eds.), *Proceedings of the 1st conference on fairness, accountability and transparency* (pp. 77–91). PMLR. <http://proceedings.mlr.press/v81/buolamwini18a.html>
- Caton, S., & Haas, C. (2020). *Fairness in machine learning: a survey a preprint*. <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>
- Celis, L. E., Huang, L., Keswani, V., & Vishnoi, N. K. (2020). Classification with fairness constraints: a meta-algorithm with provable guarantees.
- Celis, L. E., Huang, L., Keswani, V., & Vishnoi, N. K. (2021). Fair classification with noisy protected attributes: a framework with provable guarantees.
- Chen, J., Kallus, N., Mao, X., Svacha, G., & Udell, M. (2019). Fairness under unawareness. *Proceedings of the Conference on Fairness, Accountability, and Transparency*. <https://doi.org/10.1145/3287560.3287594>

- Choirat, C., Krichane, S., & Mazel-Cabasse, C. J. S. (2020). *Data in context: critical data studies i - course book*. http://isa.epfl.ch/imoniteur_ISAP/litffichecours.htm?ww_i_matiere=2708247188&ww_x_anneeAcad=1866895046&ww_i_section=116995070&ww_i_niveau=2936286&ww_c_langue=en
- Chouldechova, A. (2016). Fair prediction with disparate impact: a study of bias in recidivism prediction instruments.
- Chouldechova, A., Benavides-Prado, D., Fialko, O., & Vaithianathan, R. (2018). A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In S. A. Friedler & C. Wilson (Eds.), *Proceedings of the 1st conference on fairness, accountability and transparency* (pp. 134–148). PMLR. <http://proceedings.mlr.press/v81/chouldechova18a.html>
- College, H. (2017). *Embedded ethics*. <https://embeddedethics.seas.harvard.edu/>
- Commission, E. (2021). *Proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts*. European Commission. <https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence>
- Coughlan, S. (2019). *Oxford university promises 25% of places to disadvantaged - bbc news*. <https://www.bbc.com/news/education-48336059>
- Cramer, H., Garcia-Gathright, J., Reddy, S., Springer, A., & Takeo Bouyer, R. (2019). Translation, tracks & data: an algorithmic bias effort in practice. *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–8. <https://doi.org/10.1145/3290607.3299057>
- Crawford, K. (2013). *The hidden biases in big data*. <https://hbr.org/2013/04/the-hidden-biases-in-big-data>
- Crawford, K. (2021). *The atlas of ai: power, politics, and the planetary costs of artificial intelligence*. Yale University Press. <http://www.jstor.org/stable/j.ctv1ghv45t>
- Crenshaw, K. (1989). Demarginalizing the intersection of race and sex: a black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *The University of Chicago Legal Forum*, 140, 139–167.
- Dastin, J. (2018). Amazon scraps secret ai recruiting tool that showed bias against women. <https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G>
- Digital, C., & Office, D. (2018). *Data ethics framework - gov.uk*. <https://www.gov.uk/government/publications/data-ethics-framework>
- D'Ignazio, C., & Klein, L. F. (2020). *Data feminism*. The MIT Press.
- DrivenData. (2019). *Deon: an ethics checklist for data scientists*. <https://deon.drivendata.org/>
- Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2011). Fairness through awareness.
- Dynamics, L. W. (2009). *Adverse impact analysis / four-fifths rule*. <https://www.prevuehr.com/resources/insights/adverse-impact-analysis-four-fifths-rule/>

- Eubanks, V. (2018). *Automating inequality: how high-tech tools profile, police, and punish the poor*. St. Martin's Press, Inc.
- Feathers, T. (2021). Major universities are using race as a “high impact predictor” of student success. <https://themarkup.org/news/2021/03/02/major-universities-are-using-race-as-a-high-impact-predictor-of-student-success>
- Fiesler, C. (2018). *Tech ethics curricula: a collection of syllabi | by casey fiesler | medium*. <https://cfiesler.medium.com/tech-ethics-curricula-a-collection-of-syllabi-3eedfb76be18>
- Fiesler, C., Garrett, N., & Beard, N. (2020). What do we teach when we teach tech ethics? a syllabi analysis. *Proceedings of the 51st ACM Technical Symposium on Computer Science Education*, 289–295. <https://doi.org/10.1145/3328778.3366825>
- for Government Excellence, J. H. C. (2019). *Ethics & algorithms toolkit*. <http://ethicstoolkit.ai/>
- Foulds, J., Islam, R., Keya, K. N., & Pan, S. (2019). An intersectional definition of fairness.
- Friedler, S. A., Scheidegger, C., & Venkatasubramanian, S. (2016). On the (im)possibility of fairness. <http://arxiv.org/abs/1609.07236>
- Garage, M. I. (2019). *Ai ethics framework | digital catapult*. <https://www.migarage.ai/ethics/ethics-framework/>
- Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé, H., & Crawford, K. (2018). Datasheets for datasets. <http://arxiv.org/abs/1803.09010>
- Gilpin, L. (2018). *Mit ai ethics - building ai ethics education into computer science classes*. <https://mitaiethics.github.io/2018/11/07/ethics-education/>
- GOFF, E. L. (1983). Justice as fairness: the practice of social science in a rawlsian model. *Social Research*, 50(1), 81–97. <http://www.jstor.org/stable/40958869>
- Grosz, B. J., Grant, D. G., Vredenburg, K., Behrends, J., Hu, L., Simmons, A., & Waldo, J. (2019). Embedded ethics: integrating ethics across cs education. *Commun. ACM*, 62(8), 54–61. <https://doi.org/10.1145/3330794>
- Group, E. U. H.-I. E. (2019). *Ethics guidelines for trustworthy ai | shaping europe's digital future*. <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>
- Group, U. N. S. D. (n.d.). *Unsdg | human rights-based approach*. <https://unsdg.un.org/2030-agenda/universal-values/human-rights-based-approach>
- Hansson, J. H. (2019). About the apple card (jhh). <https://dhh.dk/2019/about-the-apple-card.html>
- Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning.
- Harwell, D. (2019). A face-scanning algorithm increasingly decides whether you deserve the job. <https://www.washingtonpost.com/technology/2019/10/22/ai-hiring-face-scanning-algorithm-increasingly-decides-whether-you-deserve-job>
- Hofmann, D. H. (1994). UCI machine learning repository. [https://archive.ics.uci.edu/ml/datasets/statlog+\(german+credit+data\)](https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data))
- Holland, S., Hosny, A., Newman, S., Joseph, J., & Chmielinski, K. (2018). The dataset nutrition label: a framework to drive higher data quality standards.
- Holstein, K., Wortman Vaughan, J., Daumé, H., Dudik, M., & Wallach, H. (2019). Improving fairness in machine learning systems: what do industry practitioners need? *Proceedings of*

- the 2019 chi conference on human factors in computing systems* (pp. 1–16). Association for Computing Machinery. <https://doi.org/10.1145/3290605.3300830>
- Hu, L., & Chen, Y. (2018). A short-term intervention for long-term fairness in the labor market. *Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW '18*. <https://doi.org/10.1145/3178876.3186044>
- Huang, G. B., Mattar, M. A., Berg, T. L., & Learned-Miller, E. (2008). Labeled faces in the wild: a database for studying face recognition in unconstrained environments.
- Iliadis, A., & Russo, F. (2016). Critical data studies: an introduction. *Big Data & Society*, 3(2), 2053951716674238. <https://doi.org/10.1177/2053951716674238>
- Jaquette, O., & Salazar, K. (2018). Colleges recruit at richer, whiter high schools. <https://www.nytimes.com/interactive/2018/04/13/opinion/college-recruitment-rich-white.html>
- Johnson, K. (2019). *How ai companies can avoid ethics washing* | *venturebeat*. <https://venturebeat.com/2019/07/17/how-ai-companies-can-avoid-ethics-washing/>
- Kamiran, F., & Calders, T. (2009). Classifying without discriminating. *2009 2nd International Conference on Computer, Control and Communication*, 1–6. <https://doi.org/10.1109/IC4.2009.4909197>
- Kamiran, F., & Calders, T. (2011). Data pre-processing techniques for classification without discrimination. *Knowledge and Information Systems*, 33. <https://doi.org/10.1007/s10115-011-0463-8>
- Kasy, M., & Abebe, R. (2021). Fairness, equality, and power in algorithmic decision-making. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 576–586. <https://doi.org/10.1145/3442188.3445919>
- Khan, F. A., Manis, E., & Stoyanovich, J. (2021). Fairness and friends. *Beyond static papers: Rethinking how we share scientific understanding in ML - ICLR 2021 workshop*. <https://openreview.net/forum?id=5PQLZP4MLEK>
- Kitchener, K. S., & Kitchener, R. F. (2009). Social science research ethics: historical and philosophical issues. *The handbook of social research ethics* (pp. 5–22). SAGE Publications, Inc. <https://doi.org/10.4135/9781483348971.n1>
- Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores.
- Krimsky, S. (2013). Do financial conflicts of interest bias research?: an inquiry into the “funding effect” hypothesis. *Science, Technology, & Human Values*, 38(4), 566–587. <https://doi.org/10.1177/0162243912456271>
- Landeau, A. (n.d.). Measuring fairness in machine learning models. <https://blog.dataiku.com/measuring-fairness-in-machine-learning-models>
- Lee, M. S. A., & Singh, J. (2021). The landscape and gaps in open source fairness toolkits. *Proceedings of the 2021 chi conference on human factors in computing systems*. Association for Computing Machinery. <https://doi.org/10.1145/3411764.3445261>
- Lehner, U. (2020). *Data beats opinion*. <https://www.letemps.ch/economie/data-beats-opinion>

- Lewis, J. R. (1992). Psychometric evaluation of the post-study system usability questionnaire: the pssuq. *Proceedings of the Human Factors Society*, 2, 1259–1263. <https://doi.org/10.1177/154193129203601617>
- Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, 22 140, 55.
- Loukides, M. K., Mason, H., & Patil, D. J. (2018). *Ethics and data science*. <http://proquest.safaribooksonline.com/?fpi=9781492043898>
- Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems 30* (pp. 4765–4774). Curran Associates, Inc. <http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf>
- Madaio, M. A., Stark, L., Wortman Vaughan, J., & Wallach, H. (2020). Co-designing checklists to understand organizational challenges and opportunities around fairness in ai. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–14. <https://doi.org/10.1145/3313831.3376445>
- Mayer, D. (2018). *Amazon killed its ai recruitment system for bias against women—report | fortune*. <https://fortune.com/2018/10/10/amazon-ai-recruitment-bias-women-sexist/>
- McGregor, L., Murray, D., & Ng, V. (2019). International human rights law as a framework for algorithmic accountability. *International and Comparative Law Quarterly*, 68(2), 309–343. <https://doi.org/10.1017/S0020589319000046>
- Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2019). A survey on bias and fairness in machine learning. <http://arxiv.org/abs/1908.09635>
- Metcalfe, J., Moss, E., & boyd danah, d. (2019). *Owning ethics: corporate logics, silicon valley, and the institutionalization of ethics* (2). <https://muse.jhu.edu/article/732185>
- Misuraca, G., & Viscusi, G. (2020). Ai-enabled innovation in the public sector: a framework for digital governance and resilience. In G. Viale Pereira, M. Janssen, H. Lee, I. Lindgren, M. P. Rodríguez Bolívar, H. J. Scholl, & A. Zuiderwijk (Eds.), *Electronic government* (pp. 110–120). Springer International Publishing.
- Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2018). Model cards for model reporting. *FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*, 220–229. <https://doi.org/10.1145/3287560.3287596>
- Mitchell, S., Potash, E., Barocas, S., D’Amour, A., & Lum, K. (2018). *Prediction-based decisions and fairness: a catalogue of choices, assumptions, and definitions*. <https://doi.org/10.1146/annurev-statistics-042720-125902>
- Morozov, E. (2013). *To save everything, click here : the folly of technological solutionism*. PublicAffairs.
- Narayanan, A. (2018). *Translation tutorial: 21 fairness definitions and their politics*. Proceedings of the Conference on Fairness, Accountability, Transparency (FAT*).

- Nasiripour, S., & Farrell, G. (2021). Goldman cleared of bias in new york review of apple card - bloomberg. <https://www.bloomberg.com/news/articles/2021-03-23/goldman-didnt-discriminate-with-apple-card-n-y-regulator-says>
- Noble, S. U. (2018). *Algorithms of oppression: how search engines reinforce racism*. NYU Press. <http://www.jstor.org/stable/j.ctt1pwt9w5>
- Ochigame, R. (2020). The long history of algorithmic fairness. <https://phenomenalworld.org/analysis/long-history-algorithmic-fairness>
- of the United Nations High Commissioner for Human Rights, O. (2004). *Iv. general recommendations adopted by the committee on the elimination of discrimination against women thirtieth session (2004)*.
- O'Neil, C. (2016). *Weapons of math destruction: how big data increases inequality and threatens democracy* (First edition). Crown.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
- Pidoux, J. (2019). Toi et moi, une distance calculée. les pratiques de quantification algorithmiques sur tinder. *Carte d'identités. L'espace au singulier*, 249–267. 370. <http://infoscience.epfl.ch/record/283981>
- Pidoux, J., Kuntz, P., & Gatica-Perez, D. (2021). Declarative variables in online dating: a mixed-method analysis of a mimetic-distinctive mechanism. *Proceedings of the ACM on Human-Computer Interaction*, CSCW, 5(CSCW1, Article 100), 100–132. <https://doi.org/10.1145/3449174>
- Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., & Weinberger, K. Q. (2017). On fairness and calibration. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems*. Curran Associates, Inc. <https://proceedings.neurips.cc/paper/2017/file/b8b9c74ac526fffb2d39ab038d1cd7-Paper.pdf>
- Poon, M. A. (2012). What lenders see – : a history of the fair isaac scorecard. <https://escholarship.org/uc/item/7n1369x2>
- Powers, D. (2008). Evaluation: from precision, recall and f-factor to roc, informedness, markedness & correlation. *Mach. Learn. Technol.*, 2.
- Raji, I. D., Scheuerman, M. K., & Amironesei, R. (2021). You can't sit with us: exclusionary pedagogy in ai ethics education. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 515–525. <https://doi.org/10.1145/3442188.3445914>
- Raso, F., Hilligoss, H., Krishnamurthy, V., Bavitz, C., & Kim, L. Y. (2018). Artificial intelligence & human rights: opportunities & risks. *Berkman Klein Center Research Publication*, 2018. <https://doi.org/10.2139/ssrn.3259344>
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, 1135–1144.

- Rights, U. N. H. (2018). *Ohchr | what are human rights?* <https://www.ohchr.org/EN/Issues/Pages/WhatareHumanRights.aspx>
- Rodolfa, K. T., Lamba, H., & Ghani, R. (2021). Empirical observation of negligible fairness-accuracy trade-offs in machine learning for public policy.
- Rowley, M. J. (2018). How ai is powering a new wave of activism. <https://newsroom.cisco.com/feature-content?type=webcontent&articleId=1947838>
- Saleiro, P., Kuester, B., Hinkson, L., London, J., Stevens, A., Anisfeld, A., Rodolfa, K. T., & Ghani, R. (2019). Aequitas: a bias and fairness audit toolkit.
- Saltz, J., Skirpan, M., Fiesler, C., Gorelick, M., Yeh, T., Heckman, R., Dewar, N., & Beard, N. (2019). Integrating ethics within machine learning courses. *ACM Trans. Comput. Educ.*, 19(4). <https://doi.org/10.1145/3341164>
- Schaake, M. (2020). Ai's invisible hand: why democratic institutions need more access to information for accountability. <https://www.rockefellerfoundation.org/blog/ais-invisible-hand-why-democratic-institutions-need-more-access-to-information-for-accountability/>
- Shorrocks, A. F. (1980). The class of additively decomposable inequality measures. *Econometrica*, 48(3), 613–625. <http://www.jstor.org/stable/1913126>
- Sokol, K., Hepburn, A., Poyiadzi, R., Clifford, M., Santos-Rodriguez, R., & Flach, P. (2020). Fat forensics: a python toolbox for implementing and deploying fairness, accountability and transparency algorithms in predictive systems. *Journal of Open Source Software*, 5(49), 1904. <https://doi.org/10.21105/joss.01904>
- Speicher, T., Heidari, H., Grgic-Hlaca, N., Gummadi, K. P., Singla, A., Weller, A., & Zafar, M. B. (2018). A unified approach to quantifying algorithmic unfairness. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. <https://doi.org/10.1145/3219819.3220046>
- Suresh, H., & Guttag, J. V. (2020). A framework for understanding unintended consequences of machine learning.
- Szczepański, M. (2019). Economic impacts of artificial intelligence (ai). *EPRS | European Parliamentary Research Service*. [https://www.europarl.europa.eu/RegData/etudes/BRIE/2019/637967/EPRS_BRI\(2019\)637967_EN.pdf](https://www.europarl.europa.eu/RegData/etudes/BRIE/2019/637967/EPRS_BRI(2019)637967_EN.pdf)
- The United Nations. (1948). *Universal declaration of human rights*.
- The United Nations. (1966). International convention on the elimination of all forms of racial discrimination. *Treaty Series*, 660, 195.
- The United Nations. (1988). Convention on the elimination of all forms of discrimination against women. *Treaty Series*, 1249, 13.
- The United Nations. (1989). Convention on the rights of the child. *Treaty Series*, 1577, 3.
- The United Nations. (2006). Convention on the rights of persons with disabilities. *Treaty Series*, 2515, 3.
- The United Nations General Assembly. (1966a). International covenant on civil and political rights. *Treaty Series*, 999, 171.
- The United Nations General Assembly. (1966b). International covenant on economic, social, and cultural rights. *Treaty Series*, 999, 171.

- UN Office of the High Commissioner for Human Rights (OHCHR). (1996). Fact sheet no. 2 (rev.1), the international bill of human rights. <https://www.ohchr.org/documents/publications/factsheet2rev.1en.pdf>
- UN Office of the High Commissioner for Human Rights (OHCHR). (2019). Un human rights business and human rights in technology project (b-tech). https://www.ohchr.org/Documents/Issues/Business/B-Tech/B_Tech_Project_revised_scoping_final.pdf
- UN Office of the High Commissioner for Human Rights (OHCHR) - Europe Regional Office. (2008). The eu and international human rights law. https://europe.ohchr.org/Documents/Publications/EU_and_International_Law.pdf
- United Nations Human Rights Committee. (1983a). *General comment: article 20 (the rights to freedom of expression and association)*.
- United Nations Human Rights Committee. (1983b). *General comment: article 23 (the right to equal pay for equal work)*.
- United Nations Human Rights Committee. (1983c). *General comment: article 7 (freedom from discrimination)*.
- UNSDG Human Rights Working Group. (2003). The human rights based approach to development cooperation towards a common understanding among un agencies, 4. https://unsdg.un.org/sites/default/files/6959-The_Human_Rights_Based_Approach_to_Development_Cooperation_Towards_a_Common_Understanding_among_UN.pdf
- V. Groeger, L. (2017). *When the designer shows up in the design*. https://www.propublica.org/article/when-the-designer-shows-up-in-the-design?utm_campaign=sprout&utm_medium=social&utm_source=twitter&utm_content=1491307350
- Vallor, S. (2019). *An ethical toolkit for engineering / design practice*. <https://www.scu.edu/ethics-in-technology-practice/ethical-toolkit/>
- Vasudevan, S., & Kenthapadi, K. (2020). LiFT: a scalable framework for measuring fairness in ml applications. *Proceedings of the 29th ACM International Conference on Information and Knowledge Management*.
- Veale, M., Van Kleek, M., & Binns, R. (2018). Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–14. <https://doi.org/10.1145/3173574.3174014>
- Verma, S., & Rubin, J. (2018). Fairness definitions explained. *IEEE/ACM International Workshop on Software Fairness*, 18. <https://doi.org/10.1145/3194770.3194776>
- Wang, A., Narayanan, A., & Russakovsky, O. (2020). Revise: a tool for measuring and mitigating bias in visual datasets.
- Waters, A., & Miikkulainen, R. (2014). Grade: machine learning support for graduate admissions. *AI Magazine*, 35, 64–75. <https://doi.org/10.1609/aimag.v35i1.2504>
- Wexler, J., Pushkarna, M., Bolukbasi, T., Wattenberg, M., Viegas, F., & Wilson, J. (2019). The what-if tool: interactive probing of machine learning models. *IEEE Transactions on Visualization and Computer Graphics*, PP, 1–1. <https://doi.org/10.1109/TVCG.2019.2934619>

- Wick, M., panda swetasudha, s., & Tristan, J.-B. (2019). Unlocking fairness: a trade-off revisited. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems*. Curran Associates, Inc. <https://proceedings.neurips.cc/paper/2019/file/373e4c5d8edfa8b74fd4b6791d0cf6dc-Paper.pdf>
- Wilson, C., Ghosh, A., Jiang, S., Mislove, A., Baker, L., Szary, J., Trindel, K., & Polli, F. (2021). Building and auditing fair algorithms: a case study in candidate screening. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 666–677. <https://doi.org/10.1145/3442188.3445928>
- Winner, L. (1980). Do artifacts have politics? *Daedalus*, 109(1), 121–136. <http://www.jstor.org/stable/20024652>
- Zafar, M. B., Valera, I., Gomez Rodriguez, M., & Gummadi, K. P. (2017). Fairness beyond disparate treatment & disparate impact. *Proceedings of the 26th International Conference on World Wide Web*. <https://doi.org/10.1145/3038912.3052660>
- Zafar, M. B., Valera, I., Rodriguez, M. G., & Gummadi, K. P. (2017). Fairness constraints: mechanisms for fair classification.