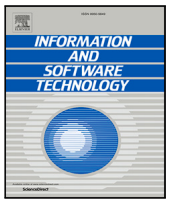
Information and Software Technology 182 (2025) 107710

Contents lists available at ScienceDirect

Information and Software Technology

journal homepage: www.elsevier.com/locate/infsof



Fairness-aware practices from developers' perspective: A survey

Gianmario Voria *, Giulia Sellitto, Carmine Ferrara, Francesco Abate, Andrea De Lucia, Filomena Ferrucci, Gemma Catolino, Fabio Palomba

Software Engineering (SeSa) Lab - Department of Computer Science, University of Salerno, Italy

ARTICLE INFO

Keywords:
Software engineering for artificial intelligence
Machine learning fairness engineering
Survey studies
Empirical software engineering

ABSTRACT

Context: Machine Learning (ML) technologies have shown great promise in many areas, but when used without proper oversight, they can produce biased results that discriminate against historically underrepresented groups. In recent years, the software engineering research community has contributed to addressing the need for ethical machine learning by proposing a number of fairness-aware practices, e.g., fair data balancing or testing approaches, that may support the management of fairness requirements throughout the software lifecycle. Nonetheless, the actual validity of these practices, in terms of practical application, impact, and effort, from the developers' perspective has not been investigated yet.

Objective: This paper addresses this limitation, assessing the developers' perspective of a set of 28 fairness practices collected from the literature.

Methods: We perform a survey study involving 155 practitioners who have been working on the development and maintenance of ML-enabled systems, analyzing the answers via statistical and clustering analysis to group fairness-aware practices based on their application frequency, impact on bias mitigation, and effort required for their application.

Results: While all the practices are deemed relevant by developers, those applied at the early stages of development appear to be the most impactful. More importantly, the effort required to implement the practices is average and sometimes high, with a subsequent average application.

Conclusion: The findings highlight the need for effort-aware automated approaches that ease the application of the available practices, as well as recommendation systems that may suggest when and how to apply fairness-aware practices throughout the software lifecycle.

1. Introduction

Artificial Intelligence (AI) and its branches, such as Machine Learning (ML), have been hot topics over the last decades, supporting humans in any decision-making activity [1] and automating repetitive tasks to reduce effort and workload [2]. ML-enabled systems, i.e., software systems that include at least one component powered by ML algorithms [3], have been deployed in several critical domains, and some recent applications in decision-making contexts such as loan management [4] or hiring decisions [5] demonstrates the potential usefulness and capabilities of such systems.

However, *every coin has a flip side*: completely relying on ML-enabled solutions without questioning them can pose risks. Many previous investigations into the ethical implications of such systems revealed how they often suffer from machine learning *fairness* concerns [6], i.e., the risk of ML models to produce outcomes that discriminate against minorities. Frequently, these issues arise from machine learning algorithms' heavy dependence on historical data, which can potentially cause them to acquire a biased understanding of the relationships governing a phenomenon. Consequently, this could result in unfair outcomes and recommendations, perpetuating discrimination and injustice against historically underrepresented groups. [7,8]. Several unfortunate cases of discrimination *caused by* ML solutions are reported in the literature, such as (1) discrimination against black people in medical cost previsioning [9], (2) biased evaluation of black people in criminal recidivism estimation [10], and (3) women discrimination in automated recruiting [11]. These examples highlight that ethical discrimination when deploying ML systems is a factual and diffused problem. It is therefore relevant for the research community to define standards to treat *ethics* and *fairness* of ML systems [12].

In response to this need, the software engineering (SE) research community, and more particularly researchers in the area of software

* Corresponding author.
E-mail addresses: gvoria@unisa.it (G. Voria), gisellitto@unisa.it (G. Sellitto), cferrara@unisa.it (C. Ferrara), f.abate20@studenti.unisa.it (F. Abate), adelucia@unisa.it (A. De Lucia), fferrucci@unisa.it (F. Ferrucci), gcatolino@unisa.it (G. Catolino), fpalomba@unisa.it (F. Palomba).

https://doi.org/10.1016/j.infsof.2025.107710
Received 31 August 2024; Received in revised form 26 February 2025; Accepted 27 February 2025
Available online 8 March 2025

engineering for artificial intelligence (SE4AI), have conducted empirical investigations and developed various guidelines, automated techniques, and tools to assist practitioners in managing fairness throughout the software lifecycle. For instance, Li et al. [13] and Chakraborty et al. [14] proposed novel approaches for fairness-aware training and hyper-parameter optimization, respectively, while Galhotra et al. [15] developed an automated method for testing software fairness properties.

Recognizing the notable advances made by researchers in recent years, our research highlights a significant limitation. Specifically, as a result of a recent mapping study [16], we defined a catalog of fairness-aware practices, i.e., common engineering practices elicited from fairness literature that have been discussed as potential bias mitigation techniques. The practices have primarily been evaluated for their potential usefulness in achieving fairness within controlled environments, such as in the case studies and experiments designed by the original authors. While fairness-aware practices are well-documented in academic literature, it remains uncertain how these practices are applied in real-world industry contexts, their perceived impact according to practitioners, and the effort required for their implementation. Gaining a deeper understanding of these aspects within industry settings is crucial for several reasons: (1) to evaluate the adoption and practical relevance of fairness-aware practices among practitioners tackling fairness-related challenges; and (2) to assess how well academic research aligns with industry needs, thereby identifying gaps that may require the development of new tools, frameworks, or perspectives to better support the integration of fairness-aware practices into real-world workflows.

This focus on industry settings is particularly relevant given the challenges in translating academic solutions into practical applications. Recent empirical findings by Friedler et al. [17] and Biswas et al. [18] highlight that the effectiveness of certain practices, i.e., their ability to increase fairness, such as fairness-aware training and optimization, may vary significantly based on specific implementation contexts. By centering our analysis on industry insights, this study aims to bridge the gap between academic research and industrial applications, offering actionable recommendations to improve the real-world adoption of fairness-aware practices.

Based on the considerations above, our study investigates the real-world applicability of fairness-aware practices systematically derived from the literature. Specifically, we aim to evaluate these practices based on their perceived impact on fairness, frequency of application, and implementation effort, as experienced by practitioners working with ML-enabled systems. More formally, we define the following research objective:

> **Our Goal**
>
> *We aim to evaluate a systematically elicited catalog of literature-derived fairness-aware practices from the practitioners' perspective, uncovering insights into their practical application, perceived impact, and the effort required for their implementation, thereby offering an industry-focused perspective on these practices.*

To achieve our objective, we conducted a survey study involving 155 practitioners working with ML-enabled software. Specifically, we gathered insights from ML practitioners, such as software engineers and data scientists, who are actively involved in developing and maintaining ML-enabled systems. This approach reflects the reality that fairness-aware practices are often implemented by general practitioners rather than specialized fairness experts [19]. We relied on our previous results of a recent mapping study [16] to select a collection of 28 fairness-aware practices. These practices were specifically tailored to address fairness concerns in ML workflows, hence focusing on mitigating bias and promoting equitable outcomes; yet, they were designed to be broadly applicable and accessible to developers, encompassing methods such as Data Balancing, Feature Transformation, and Outcome Optimization. As such, these methods align with common engineering workflows in real-world contexts while providing actionable strategies for addressing fairness challenges—this aspect further motivated the need to involve ML practitioners working on the development and maintenance of ML-enabled systems.

Practitioners were asked to rate these practices based on three indicators: *positive impact on fairness*, *frequency of application*, and *effort required for implementation*. Our findings reveal that all practices are considered to have a significant impact on fairness. However, the perceived level of effort needed for implementation varies from average to high, and the frequency of application is generally average. Based on these findings, we conclude our work by offering recommendations and insights on how to better support practitioners and provide effort-aware tools that could enhance the adoption of fairness practices in real-world contexts.

Structure of the Paper. Section 2 reports the most closely related work and motivates our study. Section 3 introduces the research questions and describe the method designed to address them. Section 4 overviews the results of our study, while Section 5 discusses the main findings and implications. Section 6 analyzes the limitations of the study and how we mitigated them. Finally, Section 7 concludes the paper and outlines our future research agenda.

2. Background and related work

This section presents a preliminary description of the problem tackled, i.e., machine learning (ML) fairness, alongside recent examples of discrimination caused by ML systems in critical contexts. Afterward, this section describes recent advances in ensuring ML fairness from different perspectives and, finally, presents the research gap that led to the definition of our research.

Background - Machine Learning Fairness. In decision-making, *fairness* is the absence of prejudice or favoritism toward an individual or group based on their inherent or acquired characteristics [6,20,21].

Fair decision-making is crucial in society, yet human bias can easily influence it [22,23]. Automated solutions like ML promise to make decisions more objective. However, recent work [12] demonstrated that sub-optimal ML development can perpetuate discrimination. This emphasizes the urgent necessity for fair ML software to prevent bias, as several cases of injustice perpetuated by AI in real-world scenarios have been reported in the literature and demonstrated the relevance of the problem in practice [24–27]. For example, criminal justice system tools like COMPAS and HART have been criticized for racially biased recidivism predictions. While COMPAS disproportionately classified Black defendants as high-risk compared to White defendants due to unbalance in the training data [10], HART's error-prone scoring raised concerns about public safety due to its misclassification of high-risk individuals as low-risk [28]. These cases highlight the need for fairness-aware practices, such as applying data mining to discover discrimination [29] in data during the Data Preparation phase or using specific validation strategies to discover biases [30] in the Model Verification & Validation step.

Financial systems also grapple with fairness issues, with models historically biased against underrepresented communities. For example, creditworthiness and lending decisions often depend on historical or proxy features, exacerbating disadvantages for Black and Hispanic individuals [31]. Practices such as diversity data selection for sensitive groups' representation [32] would help address these imbalances. Systems related to healthcare have also shown unfairness concerns, where biases in diagnostic models affect resource allocation and treatment outcomes. Discriminatory patterns emerge from under-representation in training datasets [33], such as biases in models for predicting organ transplant eligibility or cancer risk. Fairness-aware interventions, like balancing datasets used for training [34] or eliciting fair requirements

regarding all the possible underrepresented groups [35], may ensure inclusive healthcare predictions.

Other sectors highlight additional concerns. In hiring processes, AI-driven recruitment tools have displayed gender biases, favoring male candidates over female ones due to biased historical data [36]. In education, AI grading tools have been shown to unfairly disadvantage students from lower socioeconomic backgrounds due to biases in the training data [37]. Autonomous vehicles have also been found to have higher error rates in detecting pedestrians with darker skin tones [38], posing safety risks. Furthermore, language models embedded in conversational AI have been documented to produce offensive or harmful outputs, disproportionately targeting marginalized communities [39].

While these examples may not provide a comprehensive reporting of all the fairness issues arising in real-world contexts, they underline the importance of adopting fairness-aware practices across all the stages of ML development and across various critical application domains. Our work aims at addressing this need by systematically evaluating literature-derived fairness-aware practices, focusing on their application frequency, impact on bias mitigation, and the effort required for their implementation from the perspective of practitioners that regularly work on the development and maintenance with ML-enabled systems.

Related Work - Addressing Fairness. The problem of ensuring machine learning fairness has been tackled from different perspectives. Starting from the hypothesis that discrimination arises from the ML applications being trained on biased or unbalanced datasets [40], researchers have been investigating data diversity as the underlying driver of fairness. On the one hand, Zhang and Harman [41] claimed that having a dataset with many features does not help reduce discrimination; on the other hand, Chakraborty et al. [42] showed that the selection of relevant features and data heavily influences the biased outcomes. These observations pointed out that the data selection process is not trivial, and therefore needs specific attention to be properly executed. To this aim, Chakraborty et al. [42] designed *Fair-SMOTE*, a fair data balancing algorithm that does not negatively impact learning performance. Similarly, Moumoulidou et al. [43] augmented the *Max-Min* diversification objective with fairness constraints, proposing three innovative algorithms guaranteeing robust theoretical approximation tailored to varying combinations of parameters.

Drifting apart the focus on data, Brun and Meliou [12] argued that it is essential to consider fairness during the entire software life cycle. Consequently, they pointed out the need for adopting well-designed algorithms and tools to identify and report the presence of discriminatory bugs along the development pipeline. Following a similar vision, Finkelstein et al. [44] observed that several discrimination issues could reflect a poor requirement engineering process, in which the ethical aspects are formulated without considering the multi-faceted aspect of fairness. They proposed a multi-objective optimization algorithm to balance the highest number of possible fairness definitions, considering the existing conflicts that prevent fulfilling them all at once. Galhotra et al. [15] focused on the testing phase, developing a tool for the automatic generation and execution of test suites for fairness based on the principles of dataset historically underrepresented groups and data balancing. In line with the idea of considering fairness in the whole software life cycle, Richardson and Gilbert [45] explored algorithmic bias issues, fostering the creation of effective fair ML toolkits, while Caton and Haas [46] organized existing approaches and techniques to deal with fairness into *pre-processing*, *in-processing*, and *post-processing*, based on the phase they should be applied in.

Driven by the vision of bridging research and practice, Lee and Singh [47] assessed the misalignment between current open-source fairness toolkits and the practitioners' needs via exploratory focus groups, semi-structured interviews, and a survey administrated to experts. Through comparative analysis and stakeholder engagement, they identified gaps in the current toolkits' capabilities, emphasizing the need for improved support for implementing fairness in practice. Similarly, Holstein et al. [48] conducted semi-structured interviews and a survey with ML practitioners, providing a systematic dissertation of the challenges in developing fair ML systems within commercial product teams. Deng et al. [49] performed a comprehensive empirical investigation into how industry practitioners engage with existing fairness toolkits, identifying opportunities for improvement in usability and effectiveness through think-aloud interviews and an online survey. Abstracting from the technical point of view, Rakova et al. [50] investigated fairness issues from an organizational perspective. They first developed a framework to analyze how organizational culture and structure affect the effectiveness of responsible software initiatives. Through interviews with industry practitioners, they identified challenges, ethical tensions, and enablers, mapping current structures to the ideal future processes.

Research Gap. The current literature on ML fairness reports the research effort spent toward the analysis of discrimination and bias in machine learning-enabled systems, as well as the definition of novel practices or evaluation of existing ones to deal with the problem. These solutions were recently mapped from literature in a catalog of fairness-aware practices, i.e., common ML engineering practices that are developer-friendly and have been shown to improve ML fairness positively. However, to the best of our knowledge, there is still a lack of empirical evidence on the perceived impact of such approaches on fairness in practice, their frequency of adoption by practitioners, and the challenges posed by the effort required for their implementation. This gap is particularly concerning as several instances of discrimination caused by AI continue to be reported [24–27], underscoring the need for a deeper understanding of how these practices are applied and perceived in real-world contexts. In this paper, we aim to fill this gap by evaluating fairness-aware practices proposed in the literature, relying on practitioners' expertise and insights from a practical perspective.

3. Research design

We introduce our work via the Goal-Question-Metric approach [51]. The *goal* of our research was to assess literature-derived fairness-aware practices [16] from the perspective of practitioners who regularly work on ML systems. The *purpose* was to elicit insights on the perceived impact of these practices on fairness, their frequency of adoption, and the effort required for their implementation in real-world scenarios. The study serves a dual *purpose*: for researchers, it provides a better understanding of the current state of practice regarding the use of existing solutions, offering valuable input to guide the development of novel techniques and methods. For practitioners, it offers actionable insights to support more informed decision-making when adopting fairness-aware practices to address machine learning fairness concerns.

To evaluate fairness-aware practices, we defined three *qualitative* evaluation indicators, according to the software process quality theory [52]. In the first place, we were interested in assessing the quality of the practices from a practical perspective, hence assessing their actual *impact*. By impact of a practice, we mean the perceived improvement of the overall level of fairness of a machine learning solution given by the application of the specific practice. This led to the definition of the first research question of the study:

> **RQ_1: Impact**
>
> *To what extent do fairness-aware engineering practices impact bias mitigation, according to the practitioner's perspective?*

Secondly, our investigation aimed at shedding light on the extent to which fairness-aware practices are actually implemented in practice, focusing on how *frequently* practitioners adopt or are willing to adopt these practices. Hence, we asked:

RQ_2: Frequency

How frequently are fairness-aware engineering practices applied in practice, according to the practitioner's perspective?

Last but not least, we were interested in estimating the *effort* required to apply a given practice. By effort, we mean the level of complexity involved in implementing the practice within a software engineering context, i.e., how challenging the practice is to integrate into existing processes. This led to our final research question:

RQ_3: Effort

What is the effort required to apply fairness-aware engineering practices, according to the practitioner's perspective?

To address our research questions, we conducted a survey study, i.e., a structured questionnaire designed to gather insights from experienced practitioners involved in developing ML-enabled software. In doing so, we followed the guidelines by Kitchenham and Pfleeger [53], Andrews et al. [54], and Wohlin et al. [55]. In addition, we applied the *ACM/SIGSOFT Empirical Standards*;[1] in particular, we leveraged the *"General Standard"* and the *"Questionnaire Surveys"* guidelines, and the supplements on *"Sampling"* and *"Ethics"*.

All the data related to our research, i.e., the questionnaire, the complete set of anonymized answers, and the results of the analysis, are made available as part of our online appendix [56].

3.1. Context selection

The *context* of the study was represented by the set of fairness-aware practices summarized in Table 1. Specifically, the table (1) organizes the practices according to the development phases of ML-enabled systems, as outlined by Burkov [71], and (2) describes the specific actions, methodologies, and goals associated with each practice.

The practices were selected based on the findings of a *scoping review* [72], a type of secondary study that evaluates the current state of research on a specific topic by categorizing primary studies, which was conducted as part of our recent work [16]. By adhering to well-established guidelines [73], the scoping review (1) systematically identified primary studies that proposed methods for addressing fairness throughout the various development phases of ML-enabled systems and (2) organized a comprehensive catalog by grouping similar practices through iterative content analysis sessions [74].

The analysis was performed on a total of 135 articles. These articles served as the foundation for further analysis aimed at identifying fairness-aware practices. Before extracting and analyzing data to identify fairness-aware practices, we classified primary studies by paper type, research type, knowledge area, and main topic. This classification informed our data extraction and iterative content analysis, where inspectors created extraction forms, categorized practices, and refined labels until achieving theoretical saturation.

The catalog comprises 28 fairness-aware practices that are specifically tailored to address fairness concerns in ML-enabled systems. These practices span the entire lifecycle of such systems, covering the stages of design, development, maintenance, and evolution. Importantly, they build upon approaches and techniques which are likely to be already familiar to practitioners, such as data balancing, feature transformation, and model evaluation, but are adapted to explicitly mitigate bias and promote equitable outcomes. For instance, data balancing practices include techniques like oversampling underrepresented groups to reduce disparity in training data; feature transformation involves removing or masking sensitive attributes that may lead to biased predictions; and model evaluation includes fairness-specific metrics, such as demographic parity or equalized odds, to assess the fairness of ML models beyond traditional performance measures. The practices available in this catalog formed the *core object* of our survey study.

3.2. Survey structure

Fig. 1 depicts an overview of the structure of the survey. We structured the survey in four sections; we first introduced the participants to our study, then we collected information on their background; afterward, in the core section of the questionnaire, we asked them to evaluate the practices, and finally, we concluded the survey thanking them. In the following, we describe in detail the sections of the questionnaire as introduced above.

Introduction. On the welcome page of the survey, we first introduced ourselves and thanked the participants for their interest in our study. Afterward, we summarized the goal of the questionnaire and reported details on the privacy guarantees and data storage and treatment policy. We explicitly informed the participants that they were participating in our study voluntarily and, therefore, were completely free to leave the questionnaire compilation at any time. The introductory section of the survey included an explanation of the topic of machine learning fairness and the presentation of the COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) system [10]. We leveraged such a system as a case study to be later referenced in the core section of the questionnaire; namely, in the context of asking the participants to evaluate the practices, we provided them with examples of application in the COMPAS case, to facilitate the evaluation task. The rationale behind the selection of COMPAS as a scenario for our study is twofold. First, COMPAS is widely recognized as a benchmark example of a biased system in fairness literature [75] and is frequently used for evaluating fairness methods and conducting comparisons. As such, we anticipated that ML practitioners engaged with fairness-related issues would likely be familiar with this example, making it a relatable and relevant choice. Second, because COMPAS is extensively referenced in research proposing or evaluating bias mitigation methods, we were able to easily adapt the fairness-aware practices in our study to illustrate their application within this well-documented context. This ensured that the practices evaluated in the survey were both grounded in a real-world scenario and accessible to participants.

Participants' Background. The first page of questions in our survey was aimed at collecting background information from the participants. As the perception of fairness issues depends on the personal context of each individual, we first asked participants about their age, gender, and country of origin in order to properly contextualize our findings later on. However, questions asking for personal — and possibly sensitive — information were optional, and the respondents were free to continue compiling the questionnaire without answering such sensitive questions. It is worth noting that the question on gender was intended to inquire about gender identity rather than biological sex. While we used the terms "male" and "female", which may traditionally align with biological sex,[2] we explicitly sought for gender identity. In addition, we took several steps to ensure inclusivity in the survey design, many of which align with established guidelines. Specifically, we allowed participants to select more than one answer, made the question optional, and included an open-ended option for participants to self-describe their gender. These measures provided respondents with the flexibility to express their identity freely, reflecting our commitment to inclusivity and sensitivity in data collection.

We collected insights into the participants' professional backgrounds to trust their opinions on the practices they were called to evaluate. In

[1] https://github.com/acmsigsoft/EmpiricalStandards.

[2] The Morgan Klaus' gender guidelines: https://www.morgan-klaus.com/gender-guidelines.html.

Table 1
Catalog of fairness-aware practices.

Category	Practice
C1 - Requirement elicitation & analysis: Machine learning projects must have a well-defined goal. Based on the goal of the project, decisions must be taken considering crucial non-functional requirements, such as fairness.	1. Empirical methodologies for fair requirement elicitation and analysis [35] 2. Multi-objective optimization for fairness constraints [34] 3. Reverse engineering to elicit fairness requirement of a new system [35]
C2 - Data preparation: Before starting a machine learning project, the analysts must collect and prepare features and data. In this context, they must be careful for possible biases in these data.	1. Data balancing techniques to respect fairness constraints [34] 2. Data mining approaches to discover discrimination [29] 3. Data & feature transformation strategies under fairness constraints [57] 4. Diversity dataset selection for sensitive groups' representations [32] 5. Causal analysis approaches to identify discrimination dependencies in data [43] 6. Measurement approaches to improve data fairness under multiple quality constraints [57] 7. Multitask learning to maximize historically underrepresented groups' representativeness before training [6]
C3 - Model building: When selecting the right algorithm and building the model, practitioners must carefully consider fairness.	1. Ensemble learning strategies under different fairness definitions and constraints [58] 2. Focused learning strategies to obtain discrimination-free outcomes [59] 3. Fair regularization terms according to specific fairness metrics and constraints [60] 4. Adversarial learning strategies to balance fairness in quality trade-offs of the model [61]
C4 - Model training & testing: In this phase, the model is trained on the processed data and then evaluated with different metrics.	1. Fairness hyper-parameters tuning [29] 2. Post-processing transformation to balance results among historically underrepresented groups [29] 3. Post-processing strategies to optimize fairness levels of the system [62] 4. Fair test suites generation strategies [63] 5. Mutation testing for unfairness cause detection [64] 6. Testing strategies based on correct prediction oracles [48]
C5 - Model verification & validation: After training and testing the model, specific fairness constraints and trade-offs must be verified and validated.	1. Validation strategies to detect discrimination according to different meanings of data [30] 2. Features causal dependencies analysis to remove causes of discrimination [65] 3. Model comparisons for fairness level improvement [66] 4. Definitions of fairness validation strategies among different definitions and metrics [67] 5. Formal validation strategies to evaluate and improve fairness trade-offs [68]
C6 - Model maintenance & evolution: ML models must be maintained and improved by considering fair requirements or refactored to remove biases.	1. Feature standardization to improve fairness [34] 2. Model outcomes analysis to improve fairness [69] 3. Multiple datasets analysis to improve sensitive data representativeness [70]

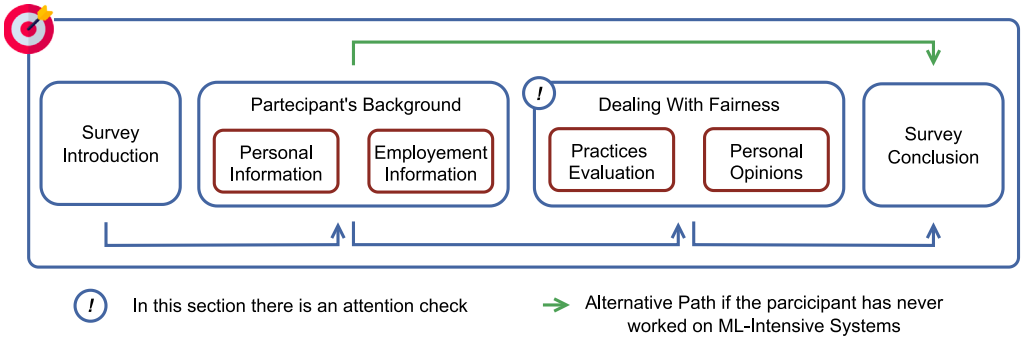


Fig. 1. Overview of the structure of the survey.

particular, we asked their level of education, their current professional role, and the number of years of experience in such role. Moreover, to ensure that participants had expertise on the topics treated in the survey, we explicitly asked them to rate their expertise on *machine learning engineering*—this allowed us to select participants matching the target audience of our study (see Section 3.4). Via an automated redirecting mechanism, those respondents who declared they had no expertise in machine learning engineering were directed to the conclusive section of the survey and thanked for their participation, therefore skipping the core task of evaluating the practices. Those who indicated familiarity and working experience in ML — even without necessarily having expertise in fairness-aware practices — were allowed to proceed to complete the questionnaire.

Practices Evaluation. In the core section of the questionnaire, we collected the participant's opinions with respect to the fairness-aware practices identified, other than additional considerations on machine learning fairness. We designed our questionnaire to meet the essential requirements of survey studies as defined in the literature [76,77]. In particular, we (1) used a clear, unambiguous, and concise vocabulary to avoid confusion among the participants, (2) preferred multiple-choice and Likert-scale questions over open-ended questions to simplify the analysis of the results, and (3) included attention checks [78] and alternative flows to identify distracted or inexpert respondents. These

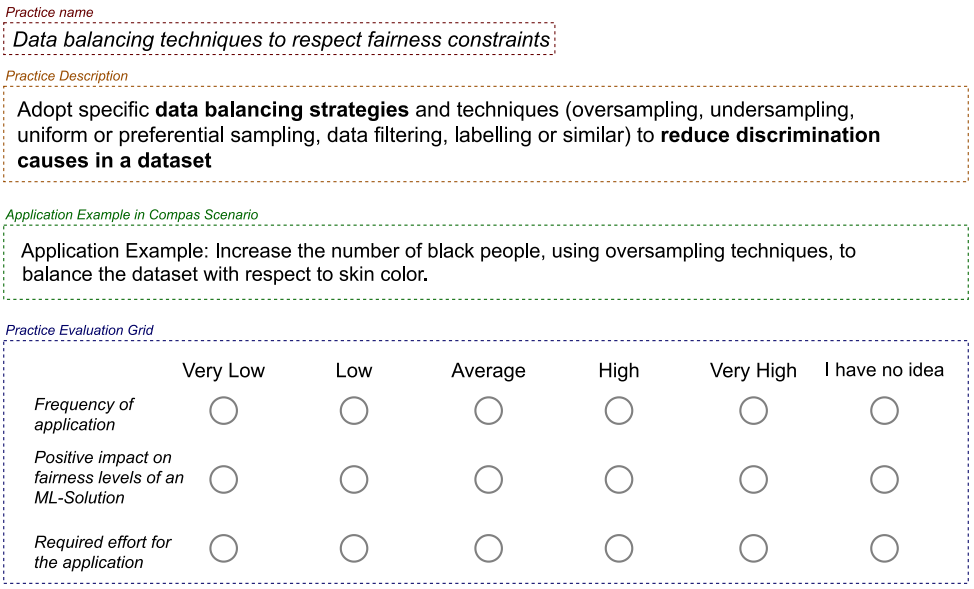


Fig. 2. Example of a fairness-aware practice evaluation grid proposed to participants.

attributes were ensured at the design time and refined through a pilot study, as discussed later in this section.

In order to gather the practitioners' opinions on the practices, we designed an *evaluation grid* that participants were asked to fill *for each* of the practices we inquired them about. To better describe the content of such a grid, we report an example in Fig. 2. First, the participants could find the practice name and description; we also included an application scenario example of the practice to make it more comprehensible, interpretable, and actionable to practitioners. To create these examples, we leveraged the authors' expertise with the elicited fairness-aware practices, instantiating each practice within the COMPAS scenario—the examples were also later validated in the context of the pilot study, as described in Section 3.3. We believe that not presenting those application examples would have made the survey excessively abstract, possibly confusing participants and affecting their judgment. We focused on COMPAS to describe the application scenarios due to its popularity; in this way, the respondents could interact with an example that they were aware of, rather than with less recognizable scenarios that would have potentially biased the evaluations and the overall conclusions of the study. We defined a set of application examples, one for each evaluated practice, tailored on COMPAS. The examples were first designed by one author of this paper, and later subject to feedback from the other authors. The complete set of application examples of the practices is available for verifiability in the online appendix of this paper [56].

The participants could rate a fairness practice employing the grid shown at the bottom of Fig. 2. The design of the grid was based on the research questions driving the study. Practitioners were asked to provide their opinion with a 5-point Likert scale ranging from *'Very low'* to *'Very high'*. As further elaborated in Section 3.4, our selection process aimed at collecting data from ML practitioners, which may and may not have expertise with the specific fairness-aware practices analyzed in our study. To account for this aspect, we allowed participants to select the *'I have no idea'* option in case they did not have enough confidence about the practice and would not like to express uninformed opinions. After the participants had assessed the entire set of practices they were requested to evaluate, they could express further opinions and feedback on the proposed practices, missed ones, and on ML fairness as a whole through a text field whose compilation was optional.

Conclusion. In the closing section of the questionnaire, we thanked the participants for taking part in our study. In addition, we provided them with an open form to share their contact information in case they were interested in receiving an overview of the preliminary results or being subsequently reached for follow-up research activities in the context of ML fairness.

3.3. Survey validation and pilot study

Before administering the survey, we validated its design by performing a pilot study using a purposive sampling strategy [79], i.e., by selecting a panel of experts in the field. We recruited three machine learning engineers in our network who have one to five years of experience in the field. When recruiting them, we refrained from providing detailed information about the specific fairness-aware practices assessed in the survey to avoid introducing potential bias. For instance, if participants had prior knowledge of the practices included in the survey, they might have unconsciously aligned their feedback to confirm or favor the practices rather than providing unbiased input on the clarity and relevance of the survey questions. Similarly, presenting detailed descriptions of fairness-aware practices upfront could have influenced their interpretations of the questions, reducing the generalizability of their feedback for practitioners with varying levels of familiarity. Instead, we focused on evaluating (1) their familiarity with common ML engineering practices and (2) the frequency with which they applied these practices in their work. This approach ensured that their feedback addressed the structure, clarity, and applicability of the survey questions without being overly influenced by the content of the specific practices.

We asked them to complete the survey and provide us with feedback on the comprehensibility of the questions, the actionability of the application examples included in the study, and the time required to complete it. In this pilot study, we also assessed the application scenarios based on COMPAS. Particularly, we asked participants to provide us with their opinions regarding (1) the clarity of the example for each fairness-aware practice and (2) the ease of understanding of the application scenario. Upon completion of the questionnaire, the pilot testers suggested various improvements to make the application scenarios clearer and more actionable, such as explicitly stating the protected attributes and the biased values assumed. In addition, they all felt that evaluating the entire catalog of practices within a reasonable time would have been impractical.

As the first step toward resolution of the raised issues, we revised the application examples according to the feedback received and re-assessed their comprehensibility, interpretability, and actionability

through a second round of interaction with the pilot testers, who were satisfied with the changes we made. In addition, we took a step further by involving a member of our research group with previous industrial experience in machine learning engineering. We asked to assess the validity of the application examples proposed in the survey. The volunteer carefully reviewed the survey and, alongside the authors, solved all the pending concerns.

We acknowledged as the second issue that the survey required too much time to be completed. Rather than administering a unique survey asking for the evaluation of all 28 fairness-aware practices, we opted for a different solution driven by well-established guidelines in survey research [80]. We split the original survey into three questionnaires, hence having smaller surveys with ten, nine, and nine fairness practices, respectively, to be evaluated. This splitting led to a more reasonable number of questions; indeed, pilot testers finally assessed the required time at 15 min, which they deemed acceptable. The sets of practices involved in each of the three split questionnaires are available in the online appendix of this paper [56]. Participants were free to answer one or more questionnaires.

3.4. Sampling strategy and survey administration

As detailed in the following paragraphs, in this stage we defined the sampling strategy, the target audience, and the procedures to administer the survey.

Sampling strategy. To select a suitable sample for our survey, we employed a cluster sampling strategy [79], a probability sampling method that divides a population into smaller, distinct groups (or clusters) and then randomly selects entire clusters to include in the sample. For our study, we turned to Prolific[3] as a platform to recruit participants. Prolific provides a balanced approach for gathering a large number of responses from practitioners with the necessary expertise, which aligns with our study's requirements. Based on prior research [81,82], we identified the Prolific community as the most appropriate cluster from which to draw our sample. To ensure the reliability and validity of the data, we selected a sample size that was suitable for the data analysis method we employed—cluster analysis. Following Hair et al.'s [83] recommendation for multivariate analysis, which suggests a minimum of five to ten observations per variable, we determined that a sample size of minimum 140 was required. This number reflects five observations for each of the 28 fairness-aware practices evaluated in the study.

Prolific is a web-based platform designed to help researchers recruit participants for survey studies. It offers the ability to apply filters to limit participation to specific, relevant subgroups. For our study, we applied the following filters: (1) Language: Fluent in English; (2) Working Sector: Science, Technology, Engineering, and Mathematics (STEM); (3) Study Level: Diploma or higher; and (4) Additional Skills: Programming skills required.

Target audience. In our study, we targeted individuals in roles related to the design, development, maintenance, and evolution of ML-enabled systems. This included (1) *'Project Managers'*, who oversee and strategically guide the implementation of ML solutions; (2) *'Software Analysts and Architects'*, whose expertise in system design and requirements elicitation helps integrate fairness considerations early in development; (3) *'Data Scientists'*, who specialize in modeling and algorithm development, and can provide valuable perspectives on fairness-aware techniques; and (4) *'Software Engineers'* and *'Data Engineers'*, whose roles are crucial in the practical implementation and deployment of ML models, often making trade-offs that can influence fairness.

The selection of the target audience was informed by multiple considerations. In the first place, we embraced the inherent challenge of gathering insights from participants having direct experience with the specific practices, prior experience with similar concepts, or theoretical knowledge. This difficulty arises in part because the population of *"machine learning fairness experts"* is not yet well-established [19]. Unlike roles such as data scientists or ML engineers, which are commonly defined in industry, a specialized role focused exclusively on fairness in ML is still emerging and not widely formalized. This observation is also supported by recent investigations into the industrial adoption of fairness tools. Nguyen et al. [84] and Holstein et al. [48] observed that fairness considerations are often a secondary priority within industrial ML workflows, embedded in roles that already focus on competing demands, such as performance optimization and scalability. In other terms, literature suggests that fairness-related tasks are performed by general practitioners, as opposed to fairness experts, during their daily tasks. As a consequence of these insights, the lack of a clearly defined population made it challenging to exclusively target practitioners with direct, extensive experience in fairness-aware practices, prompting us to involve practitioners with a broader range of expertise in ML engineering.

In addition, the fairness-aware practices evaluated in our study were derived from literature and represent methods commonly used in ML engineering workflows, but tailored to address fairness issues (see Section 3.1). This further suggests that general practitioners may be the right population to target, as the considered practices are typically applied by them. Furthermore, it is also reasonable to expect that practitioners actively involved in designing ML pipelines might have some awareness of these practices and could appropriately assess their relevance. Nonetheless, this does not mean that anyone, including practitioners with minimal experience with ML libraries, e.g., Scikit-Learn, may qualify as participants. As demonstrated by the demographics of our participants, detailed in Section 4.4, the study primarily involved technically skilled professionals with significant ML experience. This was also due to strict data analysis strategies implemented, which involved the curation of the data quality.

Despite the points above, we recognized the challenge of distinguishing between participants who are aware of a practice but do not use it and those who are entirely unfamiliar with it. In this respect, it is worth remarking that the participation in the survey was restricted to individuals who explicitly declared prior experience with ML development, ensuring that our respondents had meaningful exposure to relevant contexts. The questionnaire further emphasized that responses should be based on "working experience", and we included an "I have No Idea" option to enable participants to avoid speculating on practices they were not familiar with. By including this response option and analyzing its use across the three indicators — impact, frequency, and effort — we were able to gauge the extent to which respondents were unfamiliar with specific practices. This approach minimized the risk of collecting uninformed opinions and ensured that data from knowledgeable participants was prioritized.

Survey Administration. Reid et al. [85] have provided valuable guidelines for conducting surveys using the platform, and we made sure to follow these best practices, including pre-screening participants to ensure they met the study's criteria. This approach could ensure that our sample was both relevant and reliable, and allowed us to gather valuable insights from ML engineering practitioners in a structured and meaningful way. The three surveys were available on Prolific for a duration of 21 days, allowing sufficient time for participants to complete them.

3.5. Ethical considerations

We designed and performed our work by carefully considering the participants' privacy and the possible ethical concerns raised in survey studies [86]. We designed the survey so that all the answers were anonymous; hence, we did not collect the participants' names or email addresses. We did not ask them for any sensitive business information,

[3] Prolific website: https://www.prolific.co/.

and we explicitly guaranteed that the collected data were only used to answer our research goals. Moreover, we informed the participants that their answers would eventually be published in an aggregated form and permanently stored in the online appendix of this paper [56]. Nevertheless, gathering insights on critical aspects — such as fairness — from potential employees of organizations that could produce discriminatory ML-based products can still represent a moral concern. However, we recognize that industry practitioners have been involved in evaluating fairness and ethics in previous work [49,50]. In addition, since the scope of the survey was presented in the introduction, we believe that all the participants who answered were motivated to pursue the cause of providing non-discriminatory solutions. Still, they were free to leave the survey compilation at any time when they no longer agreed with it.

3.6. Data analysis strategy

Before addressing the three research questions, we performed a quality pre-screening of the results obtained on the general and demographic questions asked before the evaluation of the fairness-aware practices, filtering out the responses that met one or more of the following criteria: (1) lack of experience in machine learning engineering, (2) insufficient working information, (3) lack of employment information, (4) unreliable pieces of information, and (5) unavailability of answers. Regarding participants' experience in ML engineering, we directly asked whether they had prior experience with such tasks, and we excluded those who responded "No". Additionally, we omitted participants who did not provide any details about their current job or sector of employment to ensure the validity and contextual relevance of their responses. Lastly, we filtered out responses deemed unreliable, such as those where participants consistently selected "Prefer not to say" or indiscriminately chose all available options for each question. Such patterns indicate a lack of attention or engagement with the task, which could compromise the quality and integrity of the data. Note that we could not use completion time as a discriminating factor, as Prolific, the platform used to administer the survey, did not track this information. This filtering allowed us to validate the consistency and quality of the submissions and only consider valid and meaningful answers during the subsequent data analysis phase.

The data analysis strategy focused on evaluating practitioners' opinions on the impact, frequency, and effort associated with the fairness-aware practices considered. To gather a nuanced understanding of these opinions, we employed *clustering analysis* [83] for each quality indicator—impact, frequency, and effort. This method allowed us to effectively *organize* and *rank* the practices, thus providing a clearer assessment of both the ML practitioners' perspectives and the practices themselves. Indeed, by grouping similar responses, we were able to discern commonalities and variations, which helped us understand which practices were perceived as most impactful, frequently applied, or required significant effort to implement. For example, a practice being assigned to the *high impact* cluster would inform practitioners that it has a positive impact on fairness, according to the majority of the survey respondents.

From an implementation perspective, we applied cluster analysis using the Hierarchical Clustering on Principal Components (HCPC) method considering the 5-point Likert values [87,88]. Cluster analysis, also known as data segmentation, aims to group or segment a set of objects into subsets or clusters where objects within each cluster are more similar to each other than to objects in other clusters [89]. Hierarchical clustering methods divide data into clusters of varying sizes and numbers, often displaying a branching structure [90]. We applied such a method to understand meaningful patterns in the data collected, as it is suggested among the cluster analysis methods when the sample size is limited, i.e., under 200 samples [83]. We leveraged the R package *FactorMineR*[4] to compute the HCPC algorithm consisting of the following steps: (1) according to a specific evaluation criterion that depends on the type of data, the algorithm calculates the *Principal Components* of the dataset, (2) based on Ward's minimum variance [91], it computes a hierarchical clustering dendrogram based on principal components, (3) it performs an initial partition, cutting the hierarchical dendrogram, and identifying the *better number of partitions for the dataset*, and (4) finally, the algorithm executes a k-means clustering to improve the initial partition. We executed the HCPC algorithm three times, one for each research question. To highlight the clusters' characteristics and provide meaningful identifiers, we analyzed the similarity between practices within each cluster and the 5-point Likert value distributions in our catalog. This allowed us to assign summary identifiers to the clusters.

4. Analysis of the results

In this section, we present the results, starting with a preliminary analysis of sample selection and assessment. We then delve into the participants' backgrounds and address the findings for each research question, including a trade-off analysis that compares the key aspects of impact, frequency, and effort.

4.1. Preliminary sample selection

During the administration of our survey study, we collected 246 total responses, obtaining 81, 82, and 83 answers to the three questionnaires, respectively. For the sake of clarity, in the following, we report the numbers referring to each survey in parentheses, along with the total amount given by their sum.

In the pre-screening step, we removed (34+30+27=) 91 responses which were falling into one or more filtering criteria. In particular, $(27 + 24 + 20 =)$ 71 participants self-reported a lack of experience in machine learning engineering, while three respondents (one per survey) declared themselves to be students without providing sufficient information on their work experience.

We removed two participants from each sample because they did not share information about their professional roles and two submissions from each sample. After all, all multiple-choice question options in the background section were selected, suggesting that the participants were not taking the task seriously. We filtered off $(2 + 3 + 1 =)$ 6 submissions because the corresponding participants did not correctly submit the responses on the form at the end of the survey. As a consequence of this pre-screening, we obtained 47, 52, and 56 valid submissions for the first, second, and third surveys, respectively, for a total of 155 valid responses.

4.2. Preliminary sample assessment

Before applying our data analysis strategy to the collected data, we performed a preliminary assessment to mitigate a potential threat to our study. Since we gathered opinions from self-proclaimed ML engineers, for the reasons explained in Section 3, we could not be sure that they had sufficient expertise with the set of fairness-aware practices evaluated in our study. Hence, to validate our participants in terms of expertise with the practices, we analyzed the percentage of answers where they reported a *'I have No Idea'* response regarding the practice for each indicator—impact, frequency, and effort.

The lowest percentage of *'I have No Idea'* options has been achieved by the impact indicator, with a total of 60 answers spread across the 28 fairness-aware practices (4.2% of participants). Concerning the practices' frequency of application, in line with the previous indicator, a total of 62 participants answered *'I have No Idea'* (4.3% of all the

[4] https://cran.r-project.org/web/packages/FactoMineR.

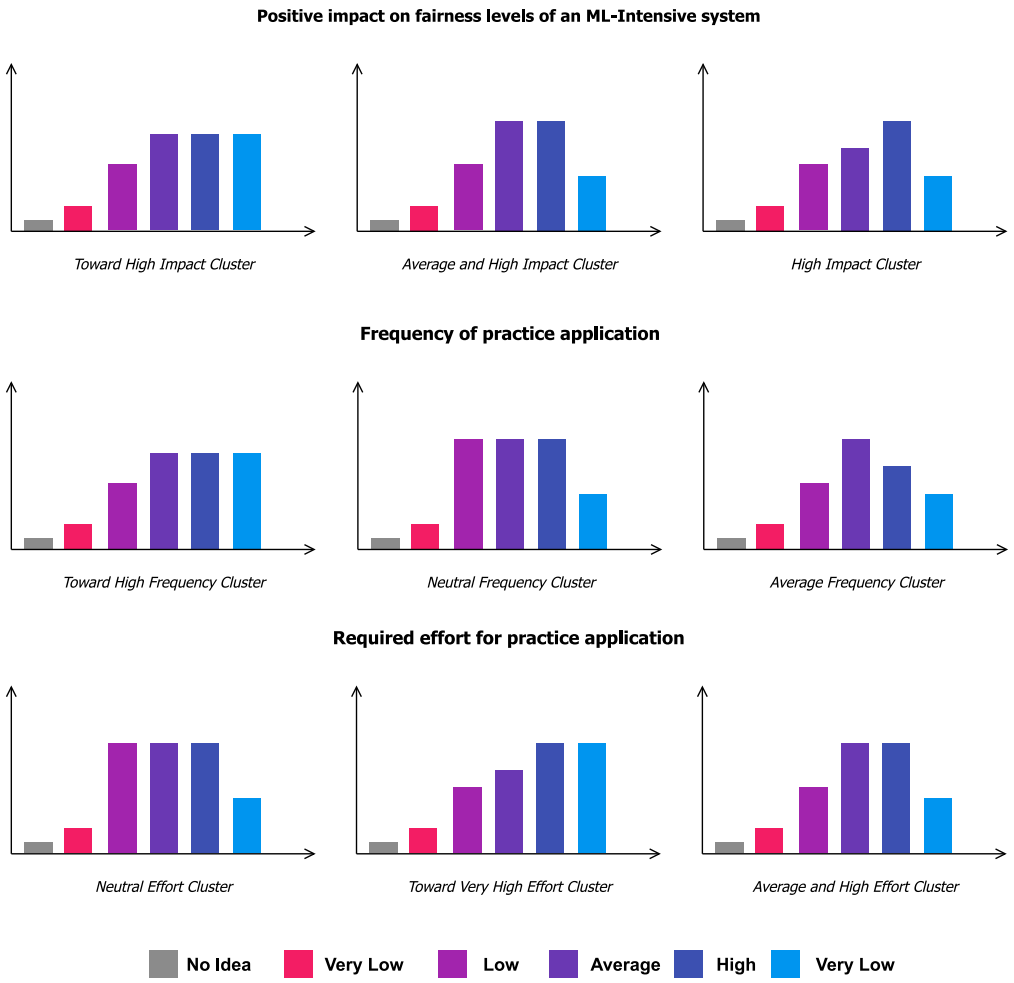


Fig. 3. Example of the typical distribution of opinions for each cluster in each indicator.

answers). Finally, 72 participants answered *'I have No Idea'* regarding the perceived effort of implementing the practices (5% of the answers).

Our analysis indicates that most of the participants were knowledgeable about the fairness-aware practices evaluated, with only a small percentage selecting *'I have No Idea'* across the three indicators. These findings reinforce the suitability of the collected data for deriving meaningful insights into fairness-aware practices, as the low levels of uncertainty minimize the risk of bias introduced by insufficient expertise. As a result, we consider the participant responses a reliable foundation for evaluating and comparing fairness-aware practices, ensuring the validity of our conclusions. Additional details about this analysis can be found in our online appendix [56].

4.3. Cluster analysis execution

After filtering the responses, we applied Cluster Analysis to provide answers to our research questions. The HCPC algorithm automatically generated three clusters for each quality attribute identified for our research questions, namely (1) the *impact* of the practice on the improvement of fairness, (2) the *frequency* of application of the practice, and (3) the *effort* required for such application. In each cluster, the practices were categorized according to the similarities in the distributions of practitioners' opinions. As a consequence, the three clusters for each quality indicator were different. Fig. 3 provides an overview of the characteristics identifying the clusters, depicting one representative example of the typical distribution of the answers collected for the practices in the same cluster for each indicator. For instance, since participants considered several practices as not frequently applied, for this indicator, the algorithm produced a *'Neutral'* cluster that groups practices for which the answers were distributed from *low* to *high*. Concerning the impact indicator, none of the practices were deemed to have a low positive impact. Hence, the lowest cluster in terms of opinions produced in this case was *'Average and High'*, which contains practices for which the answers were distributed from *average* to *high*. It is worth noting that we might have potentially forced the construction of a cluster regarding the "unknown" or "never applied" practices, i.e., those addressed by participants using the *"I have No Idea"* option. Nonetheless, as detailed in Section 4.2, the *"I have No Idea"* responses across all indicators was consistently low—below 5%. This limited proportion meant that these responses did not dominate any specific cluster, and as a result, no cluster naturally emerged to represent "unknown" or "never applied" practices in the clustering results. Fig. 6 reports the complete set of practices alongside the belonging cluster for the three quality indicators. The intensity of the colors used in such representation suggests the ranking of the cluster with respect to the others in each quality indicator. For instance, the *'Average and High'* cluster colored with lower intensity than *'Toward High'* cluster for the impact indicator suggests that the practices in the former have a lower impact than the ones in the latter.

From a practical standpoint, the clustering analysis was conducted using the HCPC algorithm implemented in the R library `FactoMine`.[5] The process was fully automated: we specified the input data, and

[5] The `FactoMine` library: https://cran.r-project.org/web/packages/FactoMineR/index.html.

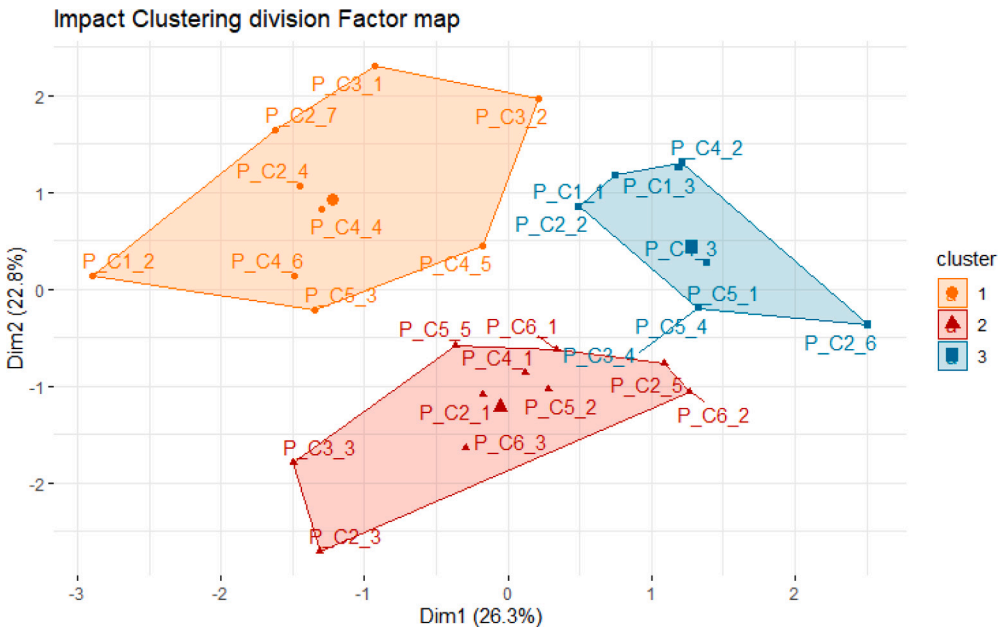


Fig. 4. Factor map generated during the HCPC analysis. This map visually represents the clustering process by showing the distribution of responses for the impact indicator across Principal Components. Practices are grouped based on similarities.

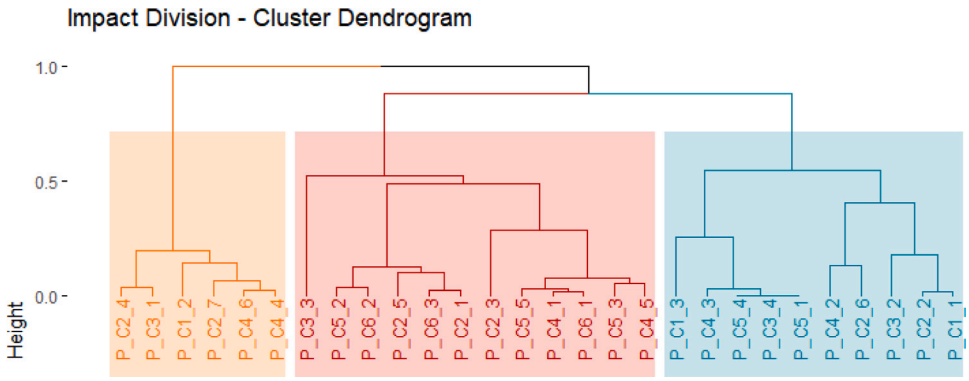


Fig. 5. Hierarchical dendrogram illustrating the clustering process for the fairness-aware practices. The dendrogram highlights the relationships between practices, showing how similar responses were grouped into clusters.

the algorithm computed the clusters based on inherent patterns in the data. The resulting clusters were then analyzed as described above, and we assigned descriptive names to each cluster (e.g., "Average Frequency" for clusters where most responses were "Average" or "High Impact" for clusters with predominantly "High" responses) to make the results interpretable and meaningful for readers. For the sake of comprehensibility and illustrating the clustering process, consider the fairness-aware practices "C1_3 - Reverse Engineering" and "C4_2 - Outcomes Transformation" for the impact indicator. These practices were both finally assigned to the "High" cluster. The responses for this practice were initially represented on a 5-point Likert scale for all the indicators, i.e., impact, frequency, and effort. In the first step, Principal Components were computed to reduce the dimensionality of the data and capture the most important variations in responses. This step grouped similar patterns of opinions while preserving as much information as possible. This is shown in Fig. 4. Subsequently, the HCPC algorithm generated a hierarchical dendrogram that highlighted relationships among responses, shown in Fig. 5. From this, an initial cluster emerged, grouping practices that showed consistent "High" responses such as "C1_3" (17 opinions) and "C4_2" (17 opinions) together. Finally, k-means clustering refined the groups, categorizing them within a cluster that was later named "High".

4.4. Participants demographic

The 155 participants in our study provide a diverse range of perspectives on the fairness-aware practices evaluated. The demographic characteristics not only offer insights into the study's participant pool but also hold implications for interpreting the results and their relevance to fairness-related challenges. As participants volunteered to participate in our study, which clearly declared to be focused on ML fairness, these data could offer insights into what kind of background people interested in ethical AI may have.

The participant pool comprises 69% men and 31% women. This gender imbalance aligns with documented trends in technology and engineering, where male professionals tend to be overrepresented [92]. As gendered experiences in the workplace often shape individuals' awareness of and engagement with fairness-related issues, the noteworthy proportion of female respondents (31%) provides valuable insights from a historically underrepresented group, which is crucial for a study centered on equity. Most of the respondents (65.8%) were aged 18–30, with 32.9% aged 31–50 and only 1.3% over 50. This distribution reflects the relatively young demographic often associated with the ML work environment. Younger professionals might be more attuned to emerging technologies and modern fairness challenges, which could influence their familiarity with and adoption of fairness-aware practices.

Practices	Impact	Frequency	Effort
C1_1 - Empirical Methodologies	High	Average	Toward Very High
C1_2 - Objective Optimization	Toward High	Neutral	Toward Very High
C1_3 - Reverse Engineering	High	Average	Neutral
C2_1 - Data Balancing	Average and High	Toward High	Neutral
C2_2 - Data Mining	High	Toward High	Average and High
C2_3 - Data and feature Transformation	Average and High	Toward High	Average and High
C2_4 - Diversity Selection	Toward High	Neutral	Neutral
C2_5 - Causal Analysis	Average and High	Average	Average and High
C2_6 - Fair Measurements	High	Neutral	Average and High
C2_7 - Multitask Learning	Toward High	Toward High	Toward Very High
C3_1 - Ensamble Learning	Toward High	Neutral	Neutral
C3_2 - Focused Learning	Toward High	Average	Neutral
C3_3 - Parameters Regularization	Average and High	Neutral	Average and High
C3_4 - Adversarial Learning	High	Average	Average and High
C4_1 - Hyperparameters Tuning	Average and High	Neutral	Average and High
C4_2 - Outcomes Transformations	High	Average	Neutral
C4_3 - Outcomes Optimization	High	Average	Neutral
C4_4 - Test Suite Generation	Toward High	Average	Neutral
C4_5 - Mutation Testing	Toward High	Neutral	Neutral
C4_6 - Testing Oracles	Toward High	Neutral	Average and High
C5_1 - Meaning Validation	High	Average	Average and High
C5_2 - Causal Validation	Average and High	Neutral	Neutral
C5_3 - Model Comparisons	Toward High	Neutral	Neutral
C5_4 - Specific Validation	High	Average	Average and High
C5_5 - Formal Validation	Average and High	Neutral	Neutral
C6_1 - Features Standardization	Average and High	Average	Neutral
C6_2 - Outcomes Analysis	Average and High	Neutral	Average and High
C6_3 - Multidataset Analysis	Average and High	Average	Average and High

Fig. 6. Practices involved in the study alongside the cluster produced for each quality indicator. The clusters are represented through a heat map according to the distribution of practitioners' opinions from the 5-point Likert scale of the questionnaire.

Participants predominantly came from Europe (74.2%), with smaller proportions from Africa (18.1%), North America (5.2%), South America (1.9%), and Asia (0.6%). This regional distribution likely reflects outreach efforts and the geographic concentration of ML practitioners. The strong European representation aligns with a regional emphasis on ethical AI and fairness in ML, as evidenced by initiatives such as the EU's AI Act. The diversity contributed by participants from Africa and the Americas could provide important contextual variations, as fairness priorities may differ depending on regional socio-economic and cultural contexts.

Furthermore, 65 participants (41.9%) identified themselves as Software Engineers and 54 participants (34.8%) as Data Scientists, making these the most represented roles. This dominance of such roles highlights the presence of participants who are technically involved in the development and deployment of ML systems. This is especially relevant, as these roles are typically tasked with managing ML workflows and may play critical parts in implementing fairness-aware practices.

Most participants reported 1–3 years of experience in their current roles (92 participants, 59.4%), followed by 4–6 years (41 participants, 26.5%) and finally, 9.7% of participants that reported more than a decade of experience. Only 2.6% of participants declared to have 0 years of experience in their current role, but they were still included in our study as they declared to have expertise with ML—that could come from their previous employment. The prominence of early-career professionals suggests that fairness concerns may resonate with individuals newer to ML-related tasks. A majority of participants hold advanced degrees: 72 (46%) have a Master's degree, and 16 (10%) possess a Ph.D. In addition, 59 participants (38%) have earned a Bachelor's degree, while very few report high school-level education. This high level of education indicates that the participants are well-prepared to engage with concepts like ML fairness, providing a robust basis for generalizing our findings to other similarly qualified ML practitioners.

The distribution of roles and educational achievements reflects a pool of respondents who are not only well-versed in ML-enabled systems but are also likely equipped to understand the challenges of fairness in AI. Moreover, the focus on early-career professionals, alongside the balanced representation of roles like Software Engineers and Data Scientists, ensures the study captures diverse perspectives without skewing toward theoretical or managerial standpoints.

4.5. RQ_1: Positive impact on fairness

The first research question of our study aimed at evaluating practitioners' opinions on the positive impact of fairness practices to assess the extent to which they believe that the practices may increase the ML fairness level of the system to which they are applied. The HCPC algorithm we executed led to the following three clusters of practices, grouped by the similarity of the answers provided by the survey respondents.

Toward High Impact. This cluster contains the nine practices for which most of the responses evaluating their impact were concentrated among the *average*, *high*, and *very high* values of the 5-point Likert scale. Most of such practices are applied during the phase of Model Training & Testing. In particular, participants considered impactful the practices of generating test suites keeping in mind the fairness requirements [63], performing mutation testing to spot possible bias [64], and providing testing oracles based on fairness [48]. As the testing phase is tightly connected to the Requirements Elicitation and Analysis one, practices

in such categories are related. Indeed, for example, participants considered that using multi-objective optimization approaches to balance different metrics and constraints in the requirements [34] has a positive impact on the fairness of the system under development.

Average and High Impact. This cluster groups the ten practices for which the *average* and *high* impact values were the maximum peaks in the distribution of the answers. All the practices in the sixth category, i.e., the ones to be applied during the stages of Model Maintenance & Evolution, fall in this cluster. In particular, analyzing the model's outcomes [69] and datasets [70] used are the practices that practitioners deemed to have an average-to-high impact on fairness. Furthermore, as leveraging feature standardization [34] falls under the broader umbrella of manipulating the models and data in general, similar practices in different categories were judged as impactful by the ML practitioners; this is the case of data balancing [34], feature transformation [57], parameters regularization [60], and hyperparameters tuning [29].

High Impact. This cluster includes the nine practices that most participants evaluated as having *high* impact on fairness. Differently from the previous two clusters, this group contains the practices on which the peak of the distribution is concentrated on the single value of *high*, highlighting the participants' unanimous opinion on their notable impact. Starting from the first phase of ML systems' life-cycle, using empirical methods such as surveys, interviews, and focus groups to elicit fairness requirements, along with applying reverse engineering on existing products [35], are highly impactful practices to be employed during Requirements Elicitation & Analysis. Afterward, in the phase of Data Preparation, using data mining approaches to find discrimination within datasets [29], and measurement optimization strategies to monitor and analyze fairness [57], are the two practices falling in the *high* impact cluster. During the stages of Model Training & Testing, and Model Verification & Validation, participants considered as having a *high* impact on fairness the practices of optimizing the outcomes of the system [62], defining custom validation strategies, and careful reasoning on socio-logical or linguistic meaning of data to detect the presence of discriminatory biases [30].

Key findings of RQ_1

In general, all the analyzed practices positively impact fairness, according to the ML practitioners involved in our study. In particular, those practices involving optimization algorithms and model adjustments were perceived to be the most impactful. Indeed, the phases of data preparation and model building are the core of ML systems' life-cycle, and the practices related to such stages were perceived to have a substantial impact on fairness, highlighting their importance in shaping equitable outcomes.

4.6. RQ_2: Frequency of application

According to practitioners, the results shed light on the varying frequencies at which different fairness practices are applied across the different stages of developing machine learning systems. In this case, the HCPC algorithm also generated three clusters of practices, as discussed below.

Toward High Frequency. This cluster includes the four practices for which the majority of responses were concentrated from the *average* to *very high* values of the 5-point Likert scale distribution. All the practices belong to the phase of Data Preparation: they consist in balancing the dataset to avoid bias [34], using data mining approaches to find discrimination within the datasets themselves [29], transforming data in order to respect focused fairness constraints [57], and performing multitask learning to maximize historically underrepresented groups' representativeness before training [6]. Such practices were judged by the ML practitioners as being the most frequently applied, highlighting the prevalence of operations related to data for the tasks of guaranteeing fairness.

Neutral Frequency. In this cluster, the algorithm included the 12 practices for which the *low*, *average*, and *high* frequency values represent comparable distribution peaks; therefore, participants did not provide consistent opinions on how frequently they are applied in practice. Such a cluster is varied and contains practices belonging to all the six phases of the ML development life-cycle. Examples of such practices are using multi-objective optimization approaches to balance different metrics and constraints [34], editing the dataset according to the similarity and diversity of data and features to improve sensitive groups' representativeness [32], and optimizing the measure of fairness before building the model [57]. In the core tasks of Model Building, Training, Testing, Verification, and Validation, several practices fall into this cluster, such as parameters regularization [60], hyperparameters tuning [29], mutation testing [64], and formal validation [68]. On the frequency of application of such practices, participants of the survey did not show a real consensus, maybe because they depend on the duration of the model's life-cycle. Therefore, we believe further work is required to understand the reasons behind this uncertainty, and to raise awareness on the importance of applying fairness-oriented practices during the whole development process of ML-enabled systems.

Average Frequency. This cluster aggregates the remaining 12 practices, for which the *average* value of frequency of application was the maximum peak in the distribution. Such practices are mostly related to the later stages of the life-cycle similarly to the previous cluster. However, while participants did not show agreement on their perspective about practices in the previous cluster, for these practices they mainly converged on the single value of an *average* frequency of application. Such practices include focused validation of the outcomes of the model [59], in particular using specific natural language processing strategies or explainable AI algorithms [30]. Furthermore, post-processing algorithms or strategies to obtain favorable outcomes for unprivileged groups [29,62] are considered applied with average frequency by the practitioners taking part in our survey study.

Key findings of RQ_2

Practices related to data preparation appear to be more commonly applied, indicating the significance of addressing fairness concerns at this stage, while practices related to model maintenance and evolution are mainly applied with average frequency. The varying application frequencies across different practices highlight the need for tailored approaches and strategies based on each ML solution's specific requirements and constraints.

On the trade-off between Impact (RQ_1) and Frequency of Application (RQ_2). Participants' responses indicated that certain fairness-aware practices, particularly those in the Data Preparation phase, e.g., data balancing and feature transformation, were implemented more frequently than others despite their perceived positive impact being around the average and high. These practices are foundational and often align closely with existing workflows. Other frequently applied practices in this stage, instead, have among the highest positive impact on fairness, according to ML practitioners, as highlighted by one of the practitioners: 💬 *I believe the pre processing methods to be far more effective (generally) than trying to address the bias during, or post, analysis*. These two insights allow us to conclude that practitioners are more engaged with fairness-aware practices that are already frequently applied in their activities, despite not being the most effective solution to mitigate bias.

Interestingly, some practices that participants ranked as having a high impact were not as frequently applied. This is particularly visible

in the later stages of development: while most practices in the stages of Model Training & Testing and Model Verification & Validation provide a high positive impact on fairness according to practitioners, they also show a low frequency of application. This disconnect between perceived impact and frequency of application highlights a potential underutilization of impactful practices due to barriers such as resource constraints, lack of awareness, or organizational support. Combining the findings of RQ_1 and RQ_2 underscores that highly impactful practices — despite their recognized value — may require greater integration and encouragement to be fully adopted in ML-intensive development.

These findings suggest the need for targeted interventions, such as dedicated training for specific practices or the development of tools to support practitioners. Future work could explore how to facilitate the broader adoption of impactful yet underutilized practices to bridge the gap between ethical needs and industry applications.

Summary

Fairness practices in the Data Preparation phase were applied most frequently and had the highest perceived impact. In addition, some practices that were deemed as frequently applied did not show the highest positive impact on fairness, perhaps indicating that practitioners are more engaged with practices that are more commonly applied in their workflows, such as data balancing, without considering the extent to which they positively impact fairness. Furthermore, impactful practices in later stages, like Model Training& Testing, were less frequently applied. This suggests barriers to adoption and highlights the need for interventions, such as targeted training or tools, to promote these practices.

4.7. RQ_3: Effort required for implementation

Our third research question was focused on assessing ML practitioners' opinions about the effort required to implement the different fairness practices during the life-cycle of ML systems. Based on what participants declared in the survey, the HCPC algorithm grouped the practices into three clusters, which we present in the following.

Neutral Effort. This cluster includes the 12 practices for which most of the opinions on their required effort were equally distributed in the *low*, *average*, and *high* values of the 5-point Likert scale in the questionnaire. Most of such practices belong to the later stages of the life-cycle, such as performing causal and formal validation of the model [65,68], and comparisons with other possible solutions, e.g., using different learning algorithms with similar parameters and configuration [66]. Similarly to what participants declared in the context of evaluating the frequency of application of such later-stage practices, they did not reach a consensus about the required effort either. This observation highlights the need for further investigation into such practices, as practitioners need to clearly know the trade-off of the required effort when deciding to apply a specific practice to their process.

Toward Very High Effort. In this cluster, the algorithm included the three practices for which *high* and *very high* values represented the maximum comparable distribution peak. Such practices all belong to the early stages of the life-cycle, i.e., requirements elicitation/analysis and data preparation. In particular, according to the practitioners' opinions, noticeable effort has to be put into the adoption of empirical strategies, such as surveys, interviews, and focus groups, to elicit and validate fairness requirements [35], and into the use of multi-objective optimization approaches to balance different constraints [34]. Furthermore, participants declared that high effort is required to apply multitasking learning strategies to maximize the average accuracy for each historically underrepresented group of the training sample [6]. All three practices in this cluster were also considered very impactful in the context of our first research questions, underlying the inevitable trade-off between the costs and benefits of applying a practice in the real world.

Average and High Effort. The algorithm included the 12 practices for which the *average* and *high* value of required effort was the maximum peak among participants' opinions. Such practices mostly belong to the phase of Data Preparation, underlying the challenge of feeding the model with proper data to guarantee that fairness requirements are met. Finding discrimination within datasets by adopting data mining techniques [29] or approaches based on causal graphs [43], manipulating data to respect fairness constraints [57], and optimizing measurement strategies [57], are practices falling into the cluster of those requiring from average to high effort. Nevertheless, practices to be applied in the core and later stages of the life-cycle are also considered to require average to high effort for their use. This is the case of parameters optimization [60] and hyperparameters tuning [29] in the core phases, and outcome analysis to be performed after deployment to evaluate and decide how to improve the fairness levels of the system in use [69].

Key findings of RQ_3

The analysis of the effort perceived by practitioners to implement the practices revealed that the early stages of development are the most effort-consuming. In particular, the stages of requirements engineering and data preparation generally require substantial effort, reflecting the intricate nature of tasks involved in these stages of ML system development. The mid stages of the ML life-cycle, particularly the phase of model training and testing, are perceived as involving more practices that require low or average effort.

On the trade-off between Impact (RQ_1) and Effort (RQ_3). Participants' responses revealed that the perceived effort associated with implementing fairness practices varied across the ML lifecycle stages. The early stages, such as Requirement Elicitation & Analysis and Data Preparation, were seen as the most effort-intensive, while later stages were perceived to require less effort. One of the practitioners involved in our survey meditated on the impact-effort trade-off by sharing their opinion: *I think using "fair" data to train the models can lead to the biggest improvement of fairness, but at the same time, it is the hardest way to solve the problem since it means throwing away lots of data, eventually already used in other experiments to which we could compare, to basically "restart from zero", leading to high times to find a working solution.*: this aligns with the resource demands of early lifecycle activities, which often involve large-scale data handling and complex pipeline setups.

However, when these findings are compared with RQ_1, an interesting trade-off emerges: participants identified several early-stage practices as both high-impact and high-effort. For instance, techniques like fairness-focused empirical methodologies for requirements elicitation [35] were deemed impactful but resource-heavy, making them a challenging yet rewarding investment. Conversely, lower-effort practices, such as fairness testing [63,64] during Model Training & Testing, offer practical entry points for organizations seeking to enhance fairness without significant resource allocation, as they appear to be among the less effort-consuming but show a toward-high positive impact on fairness.

As a general trend, practices not having the highest positive impact on fairness have been mostly retained as not very effort-consuming. This is particularly interesting when analyzed in conjunction with the frequency of application: in these cases, the frequency of application is average toward high. This may suggest that ML practitioners prioritize the application of low-effort practices despite their suboptimal impact on fairness.

These insights suggest a strategic path forward: organizations could prioritize high-impact practices early in the lifecycle while leveraging

lower-effort opportunities in later stages to maintain fairness. By aligning perceived impact with effort, it becomes possible to craft a balanced fairness strategy that optimizes resource allocation while addressing key fairness challenges.

Summary

Participants found early stages, like Data Preparation, to require more effort but offer higher impact on fairness. Later stages, such as Model Training & Testing, were seen as lower effort yet still impactful. This suggests a balanced approach, prioritizing high-impact practices early and using lower-effort ones later.

5. Discussion and implications

Our results revealed a number of discussion points, which we elaborate on in this section, along with the concrete implications that our study has on both academia and industry.

5.1. Discussion of the findings and relation to the state of the art

Our findings both align with and diverge from prior studies on fairness in ML, particularly those focused on industrial contexts. Existing research has predominantly emphasized fairness interventions during model training and dataset curation as critical stages for addressing bias [42]. While our results corroborate the importance of these stages, they also extend the discussion to other phases of the ML lifecycle that have traditionally received less attention. Specifically, practitioners in our study attributed considerable effort and impact to early-stage activities like data preparation, reinforcing the necessity of addressing fairness concerns at foundational stages of ML system development. Notably, our findings also highlight the perceived importance of underexplored lifecycle stages, such as requirements engineering and post-deployment monitoring. These phases were identified as having high potential impact on fairness but were reported to be infrequently implemented in practice. This disparity suggests that while practitioners recognize the value of these stages, barriers such as a lack of tools, frameworks, or organizational priorities may hinder their adoption. These insights align with recent calls for greater emphasis on holistic approaches to fairness that integrate considerations across the entire ML lifecycle [19]. For example, incorporating fairness-aware requirements elicitation could ensure that fairness goals are explicitly defined from the outset, while automated fairness monitoring tools could help practitioners systematically evaluate fairness outcomes post-deployment.

The evaluation of practices such as (1) constraint optimization during requirements analysis, (2) fairness assessment of data before training, (3) fairness-specific optimization post-training, and (4) targeted verification and validation strategies highlights the critical need for comprehensive fairness analytics systems. These results emphasize the importance of tools and methodologies that can effectively integrate fairness considerations across the ML lifecycle. One particularly illustrative example is the practice of analyzing model outcomes to monitor unfair behaviors, which 25 participants identified as an key strategy for enhancing fairness. This finding underscores the significance of post-training monitoring and its potential to address ongoing fairness challenges that may emerge once systems are operational. Thus, novel fairness monitoring tools hold a dual promise. First, they provide researchers with the means to conduct deeper and more systematic studies on fairness interventions, enabling the identification of nuanced biases and the development of advanced mitigation strategies. Second, they empower practitioners to implement fairness efforts in real-world development scenarios, bridging the gap between theoretical research and practical application. By enabling automated and continuous fairness assessments, these tools can serve as a cornerstone for fostering trust and accountability in AI systems across diverse application domains.

Industrial fairness research often emphasizes technical solutions, such as fairness-aware algorithms and bias mitigation techniques, as primary tools for addressing biases in AI systems [46]. While these solutions are undoubtedly important, as demonstrated in previous research on the matter, our study reveals their limitations when deployed in isolation. Human and organizational factors, such as practitioner awareness or institutional commitment, may play a pivotal role in enabling their effective adoption. This finding aligns with prior research that underscores the importance of organizational frameworks in supporting responsible AI practices [50,93]. For instance, our results indicate that practices such as fairness testing and post-prediction analysis were perceived by practitioners as highly impactful for enhancing fairness. However, their implementation in real-world scenarios remains limited, highlighting a significant gap between their perceived usefulness and practical adoption. This suggests a pressing need for fostering organizational awareness and creating structured support mechanisms to encourage fairness-oriented development. Such measures could include dedicated fairness training programs, the integration of fairness considerations into organizational workflows, and the development of collaborative tools for cross-disciplinary teams. These findings corroborate existing literature emphasizing the multifaceted nature of responsible AI practices, which require a balance between technical, human, and organizational considerations [50].

Another interesting finding from our study pertains to the perception of effort associated with fairness interventions during later lifecycle stages, such as testing, maintenance, and monitoring. According to our results, practitioners tended to rank the effort for these stages lower than traditionally reported, suggesting that they perceive fairness interventions at these stages as less complex or easier to implement. This perception contrasts with findings from industrial research, which consistently highlight inadequate attention and resource allocation to post-deployment phases as a significant challenge for ensuring fairness in AI systems [19]. While this mismatch may represent a valuable research avenue to further investigate, it suggests that redistributing fairness interventions across all stages of the ML lifecycle may offer a promising avenue for optimizing both resource allocation and fairness outcomes. More particularly, a plausible explanation of this result is that practitioners in our study, when evaluating the whole set of fairness-aware practices considered, may have recognized early-stage interventions, e.g., fairness-aware requirement elicitation and data preparation, as critical enablers for later-stage practices. In this view, a stronger emphasis on early-stage practices might reduce the perceived effort required during later stages, thereby alleviating some of the burdens traditionally associated with post-deployment fairness interventions. At the same time, enhancing post-deployment practices, i.e., automated monitoring and retraining workflows, can ensure ongoing accountability and fairness as systems evolve. Together, these strategies may address practical limitations faced by practitioners while aligning with emerging best practices for responsible AI. They emphasize fairness as a continuous, end-to-end process rather than a one-time effort, highlighting the need for sustained attention to fairness across the entire lifecycle of ML systems.

The results of our study confirm the pivotal role of training data in shaping fairness outcomes in ML systems. Data preparation, as a critical early-stage intervention, emerged as a key enabler for mitigating biases and promoting equitable results. Specifically, fairness-oriented data preparation practices, like selecting diverse datasets to ensure representation across sensitive groups and performing causal analyses to uncover relationships between features and outcomes, are gaining attention in the literature [94]. Participant evaluations in our study highlighted these practices as having significant potential to positively

impact fairness outcomes while requiring manageable effort to implement. This balance between impact and effort makes these practices particularly promising for widespread adoption in industry settings. Moreover, the growing prominence of these methods suggests they could address foundational fairness issues at the source, providing a solid groundwork for downstream interventions. These practices offer compelling opportunities for further research. Comparative studies aimed at evaluating the effectiveness of various fairness-driven data preparation techniques could provide deeper insights into their practical utility. Such studies might investigate the trade-offs between different strategies, explore their scalability to large datasets, or analyze their adaptability to domain-specific fairness challenges. By advancing our understanding of fairness in data preparation, researchers and practitioners alike can drive the development of more inclusive, robust, and impactful ML systems.

Ethical considerations in ML-enabled solutions must be closely tied to the specific application context, as fairness challenges often vary significantly across domains. Participants in our study particularly emphasized the importance of fairness practices initiated during the requirements elicitation phase. High-potential practices, such as empirical requirements elicitation and validation strategies, were identified as critical for defining fairness expectations early in the development lifecycle. However, these methods were also noted to be labor-intensive, requiring substantial time and resources to execute effectively. This finding seems to highlight the pressing need for automated or semi-automated approaches to assist practitioners in performing these tasks efficiently without compromising fairness standards. One participant reflected on the broader challenges of fairness in ML development: *"I believe some cases require building models from scratch to address application-specific fairness issues, combining various methods to achieve optimal outcomes. Unfortunately, this is often at odds with business priorities, where quick solutions are preferred. This tension is a battle I imagine other practitioners face"*. This reflection highlights a persistent trade-off between the potential fairness impact of interventions and the practical effort required to implement them. Such trade-offs often lead to prioritizing low-effort, rapid solutions over more comprehensive fairness strategies, particularly in business-driven environments. Our findings emphasize this trade-off as a defining factor in the real-world adoption of fairness-oriented techniques. Practices such as hyperparameter tuning and regularization represent low-effort strategies that are often favored due to their minimal resource demands, even though their fairness impact may be limited. Conversely, advanced practices like adversarial learning and post-processing transformations, while potentially offering higher fairness impact, require greater effort and expertise, posing barriers to adoption. Future research could investigate this trade-off, exploring the relative fairness impact and resource demands of various techniques across different application contexts. Understanding these dynamics would provide valuable insights into optimizing fairness interventions, enabling practitioners to make informed decisions based on both practical constraints and ethical imperatives. Additionally, efforts to bridge this gap through automation, tool support, or enhanced practitioner training could significantly enhance the adoption and effectiveness of fairness-aware practices across industries.

Finally, post-training evaluation practices, such as fairness testing, present promising avenues for further inquiry. Comparative studies could investigate fair mutation testing strategies or unbiased fairness oracles to assess their effectiveness in identifying fairness issues in ML systems after training, building a deeper understanding of the lifecycle dynamics of fairness interventions.

Contextualizing our Findings with Previous Research

Our study expands the scope of fairness in ML systems by addressing underexplored phases like data preparation, requirements elicitation, and post-deployment monitoring, complementing the focus on training and dataset curation found in Chakraborty et al. [42]. While aligned with Ferrara et al. [19] in advocating lifecycle approaches, we identify key barriers — such as limited tools and organizational priorities — that hinder adoption. We build on prior works [50,93] by emphasizing the integration of fairness practices, such as training and systematic monitoring, into workflows from an organizational standpoint. Specifically, we highlight fairness-specific interventions in data preparation and requirements engineering as critical enablers for downstream fairness outcomes. Our findings echo existing challenges in the industry, like balancing fairness goals with business priorities, particularly for resource-intensive interventions such as testing [84]. By proposing future work on fairness monitoring tools and comparative lifecycle evaluations, our research bridges theoretical frameworks with actionable practices, underscoring fairness as an end-to-end process.

5.2. Implications of our study

Our study explored the assessment of fairness-aware practices from the developer's perspective, highlighting several key areas for future research and actionable strategies for practitioners. In the following, we outline the implications of our findings, providing future research directions, and actionable items, considering both industry and academia perspectives. In the

5.2.1. Industry implications

Ensuring fairness in ML-enabled systems demands a comprehensive, lifecycle-wide approach, as fairness concerns permeate all stages of the development process. Our findings highlight that while the data preparation phase plays a pivotal role, other stages, such as requirements analysis, model training, validation, and verification, also contribute significantly to fairness outcomes. These insights highlight the need for industry practitioners to move beyond siloed interventions and adopt holistic strategies to address fairness challenges effectively. Below, we outline the key implications from our study that industry should consider.

Integrating Fairness Throughout Development. Fairness must be embedded into every stage of the ML development lifecycle, from requirements elicitation to post-deployment monitoring. A holistic integration ensures fairness is not treated as an isolated or secondary concern but as a fundamental pillar of the development process. One practical way to achieve this is by incorporating fairness monitoring and bias mitigation strategies into MLOps frameworks. These frameworks enable continuous assessment and refinement of fairness metrics during iterative development cycles. Automated tools, such as fairness dashboards, pipeline-integrated auditing tools, or algorithmic bias detection systems, can proactively identify and address fairness concerns in real-time. By seamlessly integrating fairness assessments into existing workflows, organizations can ensure ongoing oversight without significantly disrupting operational efficiency. Moreover, continuous fairness monitoring facilitates trust and accountability, particularly in high-stakes domains like healthcare, finance, and criminal justice.

Take-Away Message: *Integrating fairness into every stage of the ML lifecycle, supported by automated tools and MLOps frameworks, may ensure continuous oversight, builds trust, and reinforces accountability, particularly in critical application domains.*

Contextual Customization of Fairness Practices. Industry leaders must recognize that ethical considerations and fairness challenges are

highly context-dependent, varying significantly across application domains such as healthcare, finance, and criminal justice. As a result, a one-size-fits-all approach to fairness is often inadequate. Businesses must allocate resources to tailor fairness practices and requirements to their specific operational contexts and stakeholder needs. One effective strategy is *empirical requirements elicitation*, which involves gathering and validating fairness-related requirements directly from stakeholders, including end users, domain experts, and impacted communities. This approach ensures that fairness objectives are explicitly defined and aligned with the expectations and values of those most affected by the system's outcomes. Although this method may demand substantial upfront investments in time, resources, and stakeholder engagement, its long-term benefits include reducing fairness risks, enhancing system transparency, and safeguarding organizational reputation. Additionally, domain-specific customization can improve the practical applicability of fairness practices. For instance, fairness considerations in healthcare may prioritize equitable access and outcomes across demographic groups, while in finance, the focus might shift to mitigating historical biases in lending practices. By tailoring interventions to the specific domain, organizations can address fairness challenges more effectively and meaningfully. Finally, investing in contextual customization also supports compliance with evolving regulatory standards and ethical guidelines, which increasingly require organizations to demonstrate accountability for AI fairness. By adopting tailored fairness practices, businesses not only mitigate risks but also position themselves as leaders in responsible AI, fostering trust among stakeholders and competitive advantage in the marketplace.

☞ ***Take-Away Message:*** *Tailoring fairness practices to specific application contexts ensures that ethical considerations align with domain-specific challenges and stakeholder expectations. While this may require upfront investments, the long-term benefits include reduced fairness risks, enhanced transparency, and strengthened organizational reputation in an increasingly regulated and ethically conscious market.*

Balancing Effort and Impact. Fairness-enhancing practices hold undeniable potential to mitigate bias in ML systems, yet their adoption often comes with resource-intensive demands. Early-stage activities, such as fairness-aware requirements engineering and data preparation, are particularly impactful but require significant investment in time, expertise, and organizational resources. This highlights the importance of developing effort-aware solutions—strategies designed to balance the potential fairness impact of interventions with the practical constraints of implementation. Effort-aware solutions could include tools and frameworks that streamline fairness-related tasks, such as automated data auditing systems, templates for fairness-oriented requirements elicitation, and pre-configured fairness assessment pipelines. Additionally, modular fairness toolkits that allow practitioners to prioritize interventions based on available resources and organizational goals can further enhance feasibility. These tools not only make fairness initiatives more practical but also reduce the cognitive and operational burdens on development teams. By integrating such solutions into their workflows, organizations can effectively address fairness concerns without overburdening their teams or exceeding budgetary and time constraints. Moreover, this balanced approach ensures that fairness is neither overlooked due to perceived resource limitations nor implemented in a way that disrupts operational efficiency. Ultimately, effort-aware strategies provide a pragmatic pathway for achieving fairness goals while maintaining scalability and adaptability across diverse projects and industries.

☞ ***Take-Away Message:*** *Adopting effort-aware solutions enables organizations to balance the high potential impact of fairness interventions with the practical constraints of resource allocation, ensuring fairness objectives are met without overburdening development teams or disrupting workflows. This pragmatic approach supports scalable, efficient, and impactful fairness initiatives across various stages of the ML lifecycle.*

5.2.2. Academic implications

Our findings highlight the urgent need for significant advancements in research across several software engineering subfields to support fairness-aware development practices. Addressing fairness in ML systems requires innovative solutions that span the entire software engineering lifecycle, from requirements analysis to model validation and verification.

Advancing Automated Tools and Techniques. Automated tools and techniques play a pivotal role in ensuring fairness across the entire ML lifecycle, addressing both technical stages — such as model training, validation, and post-deployment monitoring — and early phases like requirements analysis. These tools can identify and mitigate biases, provide diagnostic insights, and ensure compliance with fairness metrics in real-time. However, their usefulness lies in seamless integration within MLOps frameworks, which are increasingly adopted to manage iterative development cycles in modern ML workflows. By embedding fairness monitoring and bias mitigation strategies directly into MLOps pipelines, organizations can transition from ad-hoc, manual fairness assessments to a continuous, automated approach. For example, fairness dashboards that visualize metrics like demographic parity or equalized odds can alert practitioners to emerging biases during data ingestion or model updates. Similarly, pipeline-integrated auditing tools can flag fairness risks during model retraining, enabling timely interventions without disrupting development timelines. Beyond operational benefits, automated fairness tools foster trust and accountability by providing transparent and reproducible assessments. This is especially critical in high-stakes domains such as healthcare, finance, and criminal justice, where fairness failures can have severe societal implications. Automation not only ensures consistent monitoring but also reduces the cognitive and operational burdens on practitioners, making fairness assessments scalable and sustainable across diverse application contexts.

☞ ***Take-Away Message:*** *Integrating automated fairness tools into MLOps frameworks enables real-time, continuous fairness monitoring and proactive bias correction, fostering trust, accountability, and operational efficiency across all stages of the ML lifecycle.*

Expanding Fairness Definitions Across Contexts. Fairness in machine learning is inherently context-sensitive, requiring approaches to defining and implementing fairness metrics that address the specific ethical, societal, and operational challenges of different domains. Generic fairness metrics, such as demographic parity or equalized odds, while widely adopted, often fail to capture the complex and multifaceted nature of fairness concerns unique to sectors like healthcare, criminal justice, and finance. Given the limitations of a one-size-fits-all approach, researchers may want to revisit and refine fairness definitions to ensure they resonate with domain-specific requirements. This effort should involve systematic evaluations of existing fairness metrics in diverse application scenarios. For instance, empirical studies could explore how well a particular metric performs in mitigating biases in criminal justice risk assessment tools compared to its applicability in healthcare diagnostic models. Such evaluations could uncover gaps in existing metrics, guiding the creation of new, domain-specific measures tailored to the unique demands of each field. Moreover, expanding fairness definitions across contexts should involve close collaboration with domain experts and stakeholders. Their insights can provide critical guidance on ethical priorities, practical constraints, and stakeholder expectations, ensuring that fairness interventions are grounded in real-world relevance. By incorporating these perspectives, researchers can design fairness metrics and tools that are not only theoretically robust but also practically viable and aligned with the needs of diverse industries. In addition to metric refinement, researchers should explore tools that operationalize these metrics effectively within domain-specific workflows. For instance, fairness-aware data preprocessing pipelines for healthcare applications might focus on handling imbalances in patient demographics, while finance-oriented tools could emphasize

bias correction in transactional datasets. Expanding fairness definitions to include these nuanced interventions will significantly enhance the impact and adoption of fairness practices across industries.

☞ ***Take-Away Message:*** *Refining fairness definitions and metrics to align with domain-specific requirements ensures interventions are contextually relevant, effective, and responsive to the unique challenges of diverse application settings.*

Designing Fairness-by-Design Systems. The concept of *fairness-by-design* represents a transformative shift in how fairness is approached in ML development, embedding fairness as a foundational principle rather than a reactive afterthought. These systems aim to proactively address fairness challenges at every stage of the ML lifecycle by offering context-sensitive recommendations and automated interventions tailored to the specific needs of the dataset, model, and application domain. Fairness-by-design systems leverage intelligent design principles to automate and guide fairness considerations. For example, during the data preparation phase, these tools could suggest balancing techniques to ensure equitable representation of sensitive groups, identify potential sources of bias in the dataset through causal analysis, or flag incomplete demographic attributes for further refinement. At the model training stage, they might recommend regularization methods or adversarial techniques that minimize unfair outcomes without compromising performance. Similarly, in the validation and testing stages, fairness-by-design systems could provide post-processing strategies to mitigate residual biases, ensuring compliance with ethical and regulatory standards. A key innovation of fairness-by-design systems lies in their ability to adapt dynamically to the specific requirements of different domains. These systems may use knowledge graphs or ontologies trained on domain-specific fairness guidelines, regulatory requirements, and ethical principles. For instance, a fairness-by-design tool for healthcare applications might prioritize equitable access to resources for underserved populations, while a tool for finance could emphasize reducing systemic bias in credit scoring algorithms. Moreover, fairness-by-design systems could integrate seamlessly into MLOps workflows, offering real-time recommendations and interventions as practitioners move through iterative development cycles. By automating these tasks, they reduce cognitive load on practitioners, allowing them to focus on broader system design and optimization while ensuring fairness objectives are consistently met. This integration not only enhances the efficiency of fairness interventions but also ensures they are consistently applied, reducing the likelihood of fairness violations as systems evolve.

☞ ***Take-Away Message:*** *Fairness-by-design systems proactively integrate tailored interventions across the ML lifecycle, transforming fairness from a reactive concern into a foundational principle of responsible AI development.*

Human-Centric, Effort-Aware Solutions. Fairness interventions must not only address technical challenges but also consider the human and organizational factors critical to their practical implementation. Practitioners often operate under tight deadlines, resource constraints, and varying levels of expertise, making it essential to develop fairness tools and methodologies that are both effective and user-friendly. Human-centric, effort-aware solutions aim to reduce the cognitive and operational burdens on practitioners while maintaining high standards of fairness [95]. A cornerstone of these solutions is the development of intuitive, automated tools that minimize manual effort and streamline fairness assessments. For example, tools with graphical interfaces can enable practitioners to visualize the impact of fairness interventions on datasets or models, offering actionable insights without requiring deep expertise in fairness algorithms. Similarly, pre-configured templates or workflows tailored to specific industries or regulatory requirements can reduce the complexity of implementing fairness strategies, making them accessible to a broader range of users. Cost-efficiency is another critical consideration. Many organizations, particularly smaller enterprises, may lack the resources to invest heavily in fairness-focused initiatives. Human-centric solutions should therefore be scalable and adaptable, offering functionality that aligns with diverse organizational needs and budgets. Open-source tools, modular plugins, and subscription-based platforms can provide cost-effective entry points, enabling wider adoption across the industry. Finally, human-centric solutions must prioritize accessibility and inclusivity. This includes designing interfaces that accommodate users with diverse abilities, providing multilingual support for global practitioners, and incorporating training resources to build practitioners' confidence and competence in applying fairness methodologies. By aligning tool design with the needs of end-users, these solutions can foster widespread adoption and sustainable fairness practices across diverse real-world contexts.

☞ ***Take-Away Message:*** *Human-centric, effort-aware solutions bridge the gap between fairness effectiveness and practical usability, empowering practitioners to integrate sustainable fairness practices seamlessly into their workflows.*

6. Threats to validity

Several aspects might have influenced the conclusions drawn in our survey study. Hence, in this section, we analyze the possible threats that may have biased our results.

Construct validity. The main threats to construct validity were related to the method used to identify and measure the relevance of the fairness-aware practices that practitioners evaluated. In terms of object selection, we exploited a recent scoping review [16] whose aim was to collect a comprehensive set of fairness-aware practices from the literature. These practices covered the entire lifecycle of ML-enabled systems, from design to evolution, and were identified by systematically analyzing the tools and methodologies proposed in research conducted between 2008 and 2023. By relying on the results of a scoping review featuring articles until 2023, we might have missed the analysis of some fairness-aware practices. However, we do not see this potential limitation impacting on our work since on the one hand, the broad coverage of practices provides a comprehensive foundation for understanding fairness-aware practices in ML-enabled systems, which was indeed our ultimate goal. On the other hand, the consistency observed across the practices considered suggests that any omitted practices would likely not lead to significant variations in our findings. Therefore, we believe our work makes a valuable contribution to shaping the practical relevance of fairness-aware practices in the field. In terms of reporting, we designed the survey so that each practice could be assessed through an evaluation grid: besides reporting names and descriptions of the practices, we also made sure to accompany them with actionable application scenarios revolving around the well-known COMPAS case study [10].

Concerning the use of COMPAS as a case study, we acknowledge that fairness concerns extend beyond its scope, focusing solely on tabular data classification. Although this poses a threat, we crafted the questionnaire to clarify that the case study serves as an example. Moreover, we believe that targeting experienced ML engineers partially mitigates this threat.

As for the practitioners' perception, we evaluated it based on three qualitative indicators such as impact, frequency, and effort. These metrics allowed us to have multiple views on each practice, hence making us able to provide insights into the practical relevance and usefulness of the practices. The involved practitioners were able to express their rates by adopting a 5-point Likert scale [96]. Nonetheless, we also planned the case where a practitioner could not have been confident enough of the answer and included an *'I have no idea'* option.

The three indicators also allowed us to cluster the practices by means of principal component analysis [88]. Before using clustering analysis, we verified the distribution of the answers collected and avoided misleading scale transformations.

Still, in terms of survey design, we employed well-established guidelines by Kitchenham and Pfleeger [53], Andrews et al. [54], and Wohlin

et al. [55]. In addition, we also took into account the recent considerations by Reid et al. [85], who investigated the perils of recruiting practitioners through Prolific. Specifically, we planned for the inclusion of attention checks to verify that practitioners were actually paying attention when filling out the survey.

Internal validity. In terms of internal validity, a relevant factor to consider concerns the reliability of the responses. We have employed multiple actions in this respect. First, we planned to restrict the sample allowed to participate in the survey by only considering the professional roles that might have actually provided us with reliable feedback. Through Prolific we specified filters and disclaimers that, despite being not sufficient alone, might have possibly discouraged the participation of practitioners with not enough experience/expertise on machine learning engineering. More importantly, the data quality prescreening had the goal of removing unreliable responses and, indeed, a total amount of 91 answers were deemed as invalid and removed before the analysis. In the second place, we recruited participants through Prolific, which is a platform implementing an *opt-in* strategy [97], i.e., participants volunteered to participate. This may possibly lead to self-selection or voluntary response bias, to mitigate this threat, we planned an incentive of 7 USD per valid respondent. Offering incentives has been shown to reduce bias and increase participation, as demonstrated by previous studies on survey response rates [98,99]. Moreover, the use of Prolific as a tool for gathering practitioner opinions has been positively assessed in recent research [85,100].

Another potential threat to internal validity is the limited diversity in the data collected from the sample. While we collected demographic information on age, gender, and professional role, some other demographic factors, e.g., ethnicity, sexual orientation, and cultural background, may have provided additional insights, particularly in understanding how historically underrepresented groups perceive fairness in machine learning. While this remains a potential limitation of the study, further research may be performed on this matter to complement our findings and the dimensions considered in our work.

Finally, a potential threat to internal validity arises from the varying levels of familiarity and practical experience with fairness-aware practices among participants, making it challenging to interpret the nature of our results. While fairness practices remain relatively underexplored in industry contexts [48,84,93], responses in our survey may reflect a mixture of direct application, informed theoretical perspectives, or extrapolation from general ML experiences. To mitigate this, we implemented several measures in the survey design. Participation was restricted to individuals with declared prior experience in ML development, ensuring respondents had relevant expertise. The questionnaire instructions emphasized that responses should be based on "working experience" with ML practices, encouraging participants to draw from practical knowledge rather than theoretical speculation. For participants unfamiliar with a specific practice, we included an *"I have No Idea"* option, which was used by only a few participants, increasing our confidence in the validity of the findings. Moreover, the demographics of the study revealed that participants were technically skilled professionals with significant ML experience, further supporting the reliability of their responses. While these measures aimed to ensure meaningful and realistic feedback, we acknowledge the limitations stemming from variations in expertise and the potential mix of theoretical and practical perspectives reflected in the findings.

External validity. With respect to the generalizability of the results, the majority of the survey participants declared to come from Europe. Such a higher participation from Europe was due to the distribution of users on Prolific. While we recognize this limitation, we still argue that the results reported are valuable and insightful. The empirical evaluation aimed at (1) providing an assessment of fairness-aware practices considering three qualitative indicators rather than a comprehensive and exhaustive appraisal; (2) proposing the basis for the development of further standards and empirical investigations focused on how the various practices identified could be relevant in different fair-critical contexts.

Nevertheless, future empirical studies aimed at gathering additional opinions from professional figures in other geographic areas could certainly help to obtain a more detailed empirical view with respect to the applicability of the proposed practices and their actual impact on the fairness levels of ML-enabled systems.

7. Conclusion and future work

In this paper, we evaluated 28 fairness-aware practices for improving fairness in ML-enabled systems. Through a survey involving 155 practitioners, we assessed their impact, frequency, and implementation effort. The characterization of such practices through cluster analysis provided insights that can be leveraged to understand the actual relevance of each technique better. Our findings highlight the varied challenges of mitigating bias across different stages of the ML system life cycle. Overall, all practices were deemed to impact fairness significantly. The perceived effort for implementation ranges from average to high, with the frequency of application falling around the average.

The implications of the study propose actionable items for future research in the field of software engineering for artificial intelligence, representing the input for our future research agenda. First, we aim to complement our practitioner-based analysis with software repository mining, focusing on quantifying the practices' effects on fairness metrics and understanding their adoption trends in real-world projects. Second, we plan to explore automation methods for fairness-aware practices, enabling more efficient implementation and strategies to enhance practitioners' ability to integrate fairness considerations.

Another promising direction involves defining open-ended investigations to address gaps and challenges identified in the structured evaluation of literature-derived practices. Our study highlights specific areas where academic contributions align with industry needs and areas where they fall short, presenting opportunities to explore practitioner-developed strategies that may complement or address limitations in existing approaches. Future work could employ interviews or refined surveys to examine the barriers preventing the adoption of impactful yet underutilized fairness-aware practices. Additionally, researchers might explore alternative strategies practitioners use in scenarios where predefined practices from the literature prove insufficient or impractical. Such open-ended investigations would expand the structured approach we adopted, allowing for a more holistic understanding of fairness practices in real-world settings.

CRediT authorship contribution statement

Gianmario Voria: Writing – original draft, Visualization, Validation, Investigation, Formal analysis, Data curation. **Giulia Sellitto:** Writing – original draft, Visualization, Validation, Investigation, Formal analysis, Data curation. **Carmine Ferrara:** Writing – original draft, Visualization, Validation, Investigation, Formal analysis, Data curation. **Francesco Abate:** Visualization, Validation, Investigation, Formal analysis, Data curation. **Andrea De Lucia:** Writing – review & editing, Validation, Supervision. **Filomena Ferrucci:** Writing – review & editing, Validation, Supervision. **Gemma Catolino:** Writing – review & editing, Validation, Supervision. **Fabio Palomba:** Writing – review & editing, Validation, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We acknowledge the use of ChatGPT-4 to ensure linguistic accuracy and enhance the readability of this article. This work has been partially supported by the European Union - NextGenerationEU through the Italian Ministry of University and Research, Project PRIN 2022 PNRR "FRINGE: context-aware FaiRness engineerING in complex software systEms" (grant n. P2022553SL, CUP: D53D23017340001). This work has been partially supported by the EMELIOT national research project, which has been funded by the MUR under the PRIN 2020 program (Contract 2020W3A5FY). This work was partially supported by project FAIR (PE0000013) under the NRRP MUR program funded by the EU - NGEU. We thank the reviewers and editors for their constructive comments and feedback that allowed us to significantly improve the manuscript.

Data availability

The link for publicly available data is in the article.

References

[1] Jianlong Zhou, Fang Chen, Human and machine learning, Springer, 2018.
[2] Jörg Rech, Klaus-Dieter Althoff, Artificial intelligence and software engineering: Status and future trends, KI 18 (3) (2004) 5–11.
[3] Silverio Martínez-Fernández, Justus Bogner, Xavier Franch, Marc Oriol, Julien Siebert, Adam Trendowicz, Anna Maria Vollmer, Stefan Wagner, Software engineering for AI-based systems: A survey, ACM Trans. Softw. Eng. Methodol. 31 (2) (2022) http://dx.doi.org/10.1145/3487043.
[4] Parmy Olson, The algorithm that beats your bank manager, CNN Money March 15 (2011).
[5] Claire Cain Miller, Can an algorithm hire better than a human, N. Y. Times 25 (2015).
[6] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, Aram Galstyan, A survey on bias and fairness in machine learning, ACM Comput. Surv. 54 (6) (2021) 1–35.
[7] Solon Barocas, Moritz Hardt, Arvind Narayanan, Fairness in machine learning, Nips Tutor. 1 (2017) 2.
[8] Alexandra Chouldechova, Aaron Roth, A snapshot of the frontiers of fairness in machine learning, Commun. ACM 63 (5) (2020) 82–89.
[9] Ziad Obermeyer, Brian Powers, Christine Vogeli, Sendhil Mullainathan, Dissecting racial bias in an algorithm used to manage the health of populations, Sci. 366 (6464) (2019) 447–453.
[10] Julia Angwin, Jeff Larson, Machine bias - there's software used across the country to predict future criminals. And it's biased against blacks., 2016, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. (Accessed 29 March 2022).
[11] Jeffrey Dastin, Amazon scraps secret AI recruiting tool that showed bias against women, 2018, https://cutt.ly/VKWLqF1. (Accessed 29 March 2022).
[12] Yuriy Brun, Alexandra Meliou, Proceedings of the 2018 26th acm joint meeting on european software engineering conference and symposium on the foundations of software engineering, 2018, pp. 754–759.
[13] Yanhui Li, Linghan Meng, Lin Chen, Li Yu, Di Wu, Yuming Zhou, Baowen Xu, Training data debugging for the fairness of machine learning software, in: 2022 IEEE/ACM 44th International Conference on Software Engineering, ICSE, IEEE, 2022, pp. 2215–2227.
[14] Joymallya Chakraborty, Tianpei Xia, Fahmid M Fahid, Tim Menzies, Software engineering for fairness: A case study with hyperparameter optimization, 2019, arXiv preprint arXiv:1905.05786.
[15] Sainyam Galhotra, Yuriy Brun, Alexandra Meliou, Fairness testing: testing software for discrimination, in: Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering, 2017, pp. 498–510.
[16] Gianmario Voria, Giulia Sellitto, Carmine Ferrara, Francesco Abate, Andrea De Lucia, Filomena Ferrucci, Gemma Catolino, Fabio Palomba, A catalog of fairness-aware practices in machine learning engineering, 2024, arXiv:2408.16683, URL https://arxiv.org/abs/2408.16683.
[17] Sorelle A Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P Hamilton, Derek Roth, A comparative study of fairness-enhancing interventions in machine learning, in: Proceedings of the Conference on Fairness, Accountability, and Transparency, 2019, pp. 329–338.
[18] Sumon Biswas, Hridesh Rajan, Do the machine learning models on a crowd sourced platform exhibit bias? an empirical study on model fairness, in: Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, 2020, pp. 642–653.
[19] Carmine Ferrara, Giulia Sellitto, Filomena Ferrucci, Fabio Palomba, Andrea De Lucia, Fairness-aware machine learning engineering: how far are we? Empir. Softw. Eng. 29 (1) (2024) 9.
[20] Dana Pessach, Erez Shmueli, A review on fairness in machine learning, ACM Comput. Surv. 55 (3) (2022) http://dx.doi.org/10.1145/3494672.
[21] Christopher Starke, Janine Baleis, Birte Keller, Frank Marcinkowski, Fairness perceptions of algorithmic decision-making: A systematic review of the empirical literature, Big Data Soc. 9 (2) (2022) 20539517221115189, http://dx.doi.org/10.1177/20539517221115189, arXiv:https://doi.org/10.1177/20539517221115189.
[22] Itiel E. Dror, Peter A.F. Fraser-Mackenzie, Cognitive biases in human perception, judgment, and decision making: Bridging theory and the real world, in: Criminal Investigative Failures, Routledge, 2008, pp. 79–94.
[23] Wim De Neys, Oshin Vartanian, Vinod Goel, Smarter than we think: When our brains detect that we are biased, Psychol. Sci. 19 (5) (2008) 483–489, http://dx.doi.org/10.1111/j.1467-9280.2008.02113.x, arXiv:https://doi.org/10.1111/j.1467-9280.2008.02113.x, PMID: 18466410.
[24] Tahsin Alamgir Kheya, Mohamed Reda Bouadjenek, Sunil Aryal, The pursuit of fairness in artificial intelligence models: A survey, 2024, arXiv:2403.17333, URL https://arxiv.org/abs/2403.17333.
[25] Ryan Mac, Facebook apologizes after a.I. Puts 'primates' label on video of black men, N. Y. Times (2021) URL https://www.nytimes.com/2021/09/03/technology/facebook-ai-race-primates.html.
[26] Bobbie Johnson, Helen Pidd, 'Gay writing' falls foul of amazon sales ranking system, Guardian (2009) URL https://www.theguardian.com/culture/2009/apr/13/amazon-gay-writers.
[27] Mengyi Wei, Zhixuan Zhou, AI ethics issues in real world: Evidence from AI incident database, 2023.
[28] Sheena Urwin Marion Oswald, Geoffrey C. Barnes, Algorithmic risk assessment policing models: lessons from the durham HART model and 'experimental' proportionality, Inf. Commun. Technol. Law 27 (2) (2018) 223–250, http://dx.doi.org/10.1080/13600834.2018.1458455, arXiv:https://doi.org/10.1080/13600834.2018.1458455.
[29] A. Balayn, C. Lofi, G.-J. Houben, Managing bias and unfairness in data for decision support: a survey of machine learning and data engineering approaches to identify and mitigate bias and unfairness within data management and analytics systems, VLDB J. 30 (5) (2021) 739–768, http://dx.doi.org/10.1007/s00778-021-00671-8, cited By 12.
[30] H. Bragança, J.G. Colonna, H.A.B.F. Oliveira, E. Souto, How validation methodology influences human activity recognition mobile systems, Sensors 22 (6) (2022) http://dx.doi.org/10.3390/s22062360, cited By 4.
[31] Alexander W. Butler, Erik J. Mayer, James P. Weston, Racial disparities in the auto loan market, Rev. Financ. Stud. 36 (1) (2022) 1–41, http://dx.doi.org/10.1093/rfs/hhac029, arXiv:https://academic.oup.com/rfs/article-pdf/36/1/1/48135025/hhac029.pdf.
[32] D. Wang, L. Cheng, T. Wang, Fairness-aware genetic-algorithm-based few-shot classification, Math. Biosci. Eng. 20 (2) (2023) 3624–3637, http://dx.doi.org/10.3934/mbe.2023169, cited By 0.
[33] Mildred Cho, Rising to the challenge of bias in health care AI, Nature Med. 27 (2021) 1–2, http://dx.doi.org/10.1038/s41591-021-01577-2.
[34] S. Biswas, H. Rajan, Fair preprocessing: Towards understanding compositional fairness of data transformers in machine learning pipeline, in: Spinellis D. (Ed.), ESEC/FSE 2021 - Proceedings of the 29th ACM Joint Meeting European Software Engineering Conference and Symposium on the Foundations of Software Engineering, 2021, pp. 981–993, http://dx.doi.org/10.1145/3468264.3468536, cited By 12.
[35] A. Lewis, J. Stoyanovich, Teaching responsible data science: Charting new pedagogical territory, Int. J. Artif. Intell. Educ. 32 (3) (2022) 783–807, http://dx.doi.org/10.1007/s40593-021-00241-7, cited By 3.
[36] Dena F. Mujtaba, Nihar R. Mahapatra, Ethical considerations in AI-based recruitment, in: 2019 IEEE International Symposium on Technology and Society, ISTAS, IEEE, 2019, pp. 1–7.
[37] Ryan S. Baker, Aaron Hawn, Algorithmic bias in education, Int. J. Artif. Intell. Educ. (2022) 1–41.
[38] Benjamin Wilson, Judy Hoffman, Jamie Morgenstern, Predictive inequity in object detection, 2019, arXiv preprint arXiv:1902.11097.
[39] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, Noah A Smith, RealToxicityPrompts: Evaluating neural toxic degeneration in language models, in: Findings of the Association for Computational Linguistics: EMNLP 2020, 2020, pp. 3356–3369.
[40] Sriram Vasudevan, Krishnaram Kenthapadi, Lift: A scalable framework for measuring fairness in ml applications, in: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, 2020, pp. 2773–2780.
[41] Jie M. Zhang, Mark Harman, "Ignorance and prejudice" in software fairness, in: 2021 IEEE/ACM 43rd International Conference on Software Engineering, ICSE, IEEE, 2021, pp. 1436–1447.
[42] Joymallya Chakraborty, Suvodeep Majumder, Tim Menzies, Bias in machine learning software: why? how? what to do? in: Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, 2021, pp. 429–440.

[43] Z. Moumoulidou, A. McGregor, A. Meliou, Diverse data selection under fairness constraints, in: K. Yi, Z. Wei (Eds.), Leibniz International Proceedings in Informatics, LIPIcs, Vol. 186, 2021, http://dx.doi.org/10.4230/LIPIcs.ICDT.2021.13, cited By 3.

[44] Anthony Finkelstein, Mark Harman, S Afshin Mansouri, Jian Ren, Yuanyuan Zhang, "Fairness analysis" in requirements assignments, in: 2008 16th IEEE International Requirements Engineering Conference, IEEE, 2008, pp. 115–124.

[45] Brianna Richardson, Juan E. Gilbert, A framework for fairness: A systematic review of existing fair AI solutions, 2021, CoRR abs/2112.05700. arXiv:2112.05700. URL https://arxiv.org/abs/2112.05700.

[46] Simon Caton, Christian Haas, Fairness in machine learning: A survey, ACM Comput. Surv. 56 (7) (2024) http://dx.doi.org/10.1145/3616865.

[47] Michelle Seng Ah Lee, Jat Singh, The landscape and gaps in open source fairness toolkits, in: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, CHI '21, Association for Computing Machinery, New York, NY, USA, 2021, http://dx.doi.org/10.1145/3411764.3445261.

[48] K. Holstein, J.W. Vaughan, I.I.I. Daumé, M. Dudík, H. Wallach, Improving fairness in machine learning systems: What do industry practitioners need? in: Conference on Human Factors in Computing Systems - Proceeding, 2019, http://dx.doi.org/10.1145/3290605.3300830, cited By 214.

[49] Wesley Hanwen Deng, Manish Nagireddy, Michelle Seng Ah Lee, Jatinder Singh, Zhiwei Steven Wu, Kenneth Holstein, Haiyi Zhu, Exploring how machine learning practitioners (try to) use fairness toolkits, in: 2022 ACM Conference on Fairness, Accountability, and Transparency, in: FAccT 201922, ACM, 2022, http://dx.doi.org/10.1145/3531146.3533113, URL http://dx.doi.org/10.1145/3531146.3533113.

[50] Bogdana Rakova, Jingying Yang, Henriette Cramer, Rumman Chowdhury, Where responsible AI meets reality: Practitioner perspectives on enablers for shifting organizational practices, Proc. ACM Hum.- Comput. Interact. 5 (CSCW1) (2021) http://dx.doi.org/10.1145/3449081.

[51] V.R. Basili, Software Modeling and Measurement: The Goal/Question/metric Paradigm, University of Maryland, 1992, URL https://books.google.it/books?id=Gc-cpwAACAAJ.

[52] Ron S. Kenett, Emanuel Baker, Software Process Quality: Management and Control, CRC Press, 1999.

[53] Barbara A. Kitchenham, Shari Lawrence Pfleeger, Principles of survey research part 2: designing a survey, ACM SIGSOFT Softw. Eng. Notes 27 (1) (2002) 18–20.

[54] Dorine Andrews, Blair Nonnecke, Jennifer Preece, Conducting research on the internet:: Online survey design, development and implementation guidelines, 2007.

[55] Claes Wohlin, Per Runeson, Martin Höst, Magnus C Ohlsson, Björn Regnell, Anders Wesslén, Experimentation in Software Engineering, Springer Science and Business Media, 2012.

[56] Gianmario Voria, Giulia Sellitto, Carmine Ferrara, Francesco Abate, Andrea De Lucia, Filomena Ferrucci, Gemma Catolino, Fabio Palomba, Online Appendix. URL https://figshare.com/s/67ea3abd65f91bb36379.

[57] F. Nargesian, A. Asudeh, H.V. Jagadish, Responsible data integration: Next-generation challenges, in: Proceedings of the ACM SIGMOD International Conference on Management of Data, 2022, pp. 2458–2464, http://dx.doi.org/10.1145/3514221.3522567, cited By 1.

[58] Z. Chen, J.M. Zhang, F. Sarro, M. Harman, MAAT: a novel ensemble approach to addressing fairness and performance bugs for machine learning software, in: ESEC/FSE 2022 - Proceedings of the 30th ACM Joint Meeting European Software Engineering Conference and Symposium on the Foundations of Software Engineering, 2022, pp. 1122–1134, http://dx.doi.org/10.1145/3540250.3549093, cited By 0.

[59] A.J. Biega, K.P. Gummadi, G. Weikum, Equity of attention: Amortizing individual fairness in rankings, in: 41st International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2018, 2018, pp. 405–414, http://dx.doi.org/10.1145/3209978.3210063, cited By 181.

[60] M. Vega-Gonzalo, P. Christidis, Fair models for impartial policies: Controlling algorithmic bias in transport behavioural modelling, Sustain. (Switz.) 14 (14) (2022) http://dx.doi.org/10.3390/su14148416, cited By 0.

[61] Y. Zhang, L. Luo, H. Huang, Unified fairness from data to learning algorithm, in: Proceedings - IEEE International Conference on Data Mining, ICDM 2021, Vol. 2021-December, 2021, pp. 1499–1504, http://dx.doi.org/10.1109/ICDM51629.2021.00195, cited By 1.

[62] M. Hort, J.M. Zhang, F. Sarro, M. Harman, Fairea: A model behaviour mutation approach to benchmarking bias mitigation methods, in: Spinellis D. (Ed.), ESEC/FSE 2021 - Proceedings of the 29th ACM Joint Meeting European Software Engineering Conference and Symposium on the Foundations of Software Engineering, 2021, pp. 994–1006, http://dx.doi.org/10.1145/3468264.3468565, cited By 10.

[63] P.K. Lohia, K. Natesan Ramamurthy, M. Bhide, D. Saha, K.R. Varshney, R. Puri, Bias mitigation post-processing for individual and group fairness, in: ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, Vol. 2019-May, Institute of Electrical and Electronics Engineers Inc., 2019, pp. 2847–2851, http://dx.doi.org/10.1109/ICASSP.2019.8682620, cited By 34.

[64] Z. Chen, J.M. Zhang, F. Sarro, M. Harman, MAAT: a novel ensemble approach to addressing fairness and performance bugs for machine learning software, in: Kim M. Roychoudhury A. (Ed.), ESEC/FSE 2022- Proceedings of the 30th ACM Joint Meeting European Software Engineering Conference and Symposium on the Foundations of Software Engineering, 2022, pp. 1122–1134, http://dx.doi.org/10.1145/3540250.3549093, cited By 0.

[65] J. Gardner, C. Brooks, R. Baker, Evaluating the fairness of predictive student models through slicing analysis, in: ACM International Conference Proceeding Series, 2019, pp. 225–234, http://dx.doi.org/10.1145/3303772.3303791, cited By 61.

[66] R.S. Baker, A. Hawn, Algorithmic bias in education, Int. J. Artif. Intell. Educ. 32 (4) (2022) 1052–1092, http://dx.doi.org/10.1007/s40593-021-00285-9, cited By 21.

[67] S. Raza, D.J. Reji, C. Ding, Dbias: detecting biases and ensuring fairness in news articles, Int. J. Data Sci. Anal. (2022) http://dx.doi.org/10.1007/s41060-022-00359-4, cited By 0.

[68] P. Chejara, L.P. Prieto, A. Ruiz-Calleja, M.J. Rodríguez-Triana, S.K. Shankar, R. Kasepalu, Efar-mmla: An evaluation framework to assess and report generalizability of machine learning models in mmla, Sensors 21 (8) (2021) http://dx.doi.org/10.3390/s21082863, cited By 4.

[69] Y. Yang, C. Zhang, C. Fan, A. Mostafavi, X. Hu, Towards fairness-aware disaster informatics: An interdisciplinary perspective, IEEE Access 8 (2020) 201040–201054, http://dx.doi.org/10.1109/ACCESS.2020.3035714, cited By 4.

[70] T. Zhang, T. Zhu, J. Li, M. Han, W. Zhou, P.S. Yu, Fairness in semi-supervised learning: Unlabeled data help to reduce discrimination, IEEE Trans. Knowl. Data Eng. 34 (4) (2022) 1763–1774, http://dx.doi.org/10.1109/TKDE.2020.3002567, cited By 7.

[71] Andriy Burkov, Machine Learning Engineering, vol. 1, True Positive Incorporated, 2020.

[72] Paul Ralph, Sebastian Baltes, Paving the way for mature secondary research: the seven types of literature review, in: Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering, 2022, pp. 1632–1636.

[73] Kai Petersen, Sairam Vakkalanka, Ludwik Kuzniarz, Guidelines for conducting systematic mapping studies in software engineering: An update, Inf. Softw. Technol. 64 (2015) 1–18.

[74] Marilyn Domas White, Emily E. Marsh, Content analysis: A flexible methodology, Libr. Trends 55 (1) (2006) 22–45.

[75] Alessandro Fabris, Stefano Messina, Gianmaria Silvello, Gian Antonio Susto, Algorithmic fairness datasets: the story so far, Data Min. Knowl. Discov. 36 (6) (2022) 2074–2152, http://dx.doi.org/10.1007/s10618-022-00854-z.

[76] Arlene Fink, The Survey Handbook, sage, 2003.

[77] Elizabeth A. Buchanan, Erin E. Hvizdak, Online survey tools: Ethical and methodological concerns of human research ethics committees, J. Empir. Res. Hum. Res. Ethics 4 (2) (2009) 37–48.

[78] Tobias Gummer, Joss Roßmann, Henning Silber, Using instructed response items as attention checks in web surveys: Properties and implementation, Sociol. Methods Res. 50 (1) (2021) 238–264.

[79] Sebastian Baltes, Paul Ralph, Sampling in software engineering research: A critical review and guidelines, Empir. Softw. Eng. 27 (4) (2022) 94.

[80] Carla A Green, Naihua Duan, Robert D Gibbons, Kimberly E Hoagwood, Lawrence A Palinkas, Jennifer P Wisdom, Approaches to mixed methods dissemination and implementation research: methods, strengths, caveats, and opportunities, Adm. Policy Ment. Heal. Ment. Heal. Serv. Res. 42 (5) (2015) 508–523.

[81] Daniel Russo, Navigating the complexity of generative AI adoption in software engineering—RCR report, ACM Trans. Softw. Eng. Methodol. 33 (8) (2024) http://dx.doi.org/10.1145/3680471.

[82] Stefano Lambiase, Gemma Catolino, Fabio Palomba, Filomena Ferrucci, Daniel Russo, Investigating the role of cultural values in adopting large language models for software engineering, 2024, arXiv preprint arXiv:2409.05055.

[83] Joseph Hair, William Black, Barry Babin, Rolph Anderson, Multivariate data analysis: A global perspective, ISBN: 0135153093, 2010.

[84] Thanh Nguyen, Maria Teresa Baldassarre, Luiz Fernando de Lima, Ronnie de Souza Santos, From literature to practice: Exploring fairness testing tools for the software industry adoption, in: Proceedings of the 18th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, ESEM '24, Association for Computing Machinery, New York, NY, USA, 2024, pp. 549–555, http://dx.doi.org/10.1145/3674805.3695404.

[85] Brittany Reid, Markus Wagner, Marcelo d'Amorim, Christoph Treude, Software engineering user study recruitment on prolific: An experience report, 2022, arXiv preprint arXiv:2201.05348.

[86] T. Hall, V. Flynn, Ethical issues in software engineering research: a survey of current practice, Empir. Softw. Eng. 6 (4) (2001) 305–317.

[87] Alboukadel Kassambara, Practical Guide to Principal Component Methods in R: PCA, M (CA), FAMD, MFA, HCPC, factoextra, vol. 2, Sthda, 2017.

[88] M. Argüelles, C. Benavides, I. Fernández, A new approach to the identification of regional clusters: hierarchical clustering on principal components, Appl. Econ. 46 (21) (2014) 2511–2519.

[89] Charles Romesburg, Cluster Analysis for Researchers, Lulu. com, 2004.

[90] Cecil C. Bridges Jr., Hierarchical cluster analysis, Psychol. Rep. 18 (3) (1966) 851–854.
[91] Gabor J. Szekely, Maria L. Rizzo, et al., Hierarchical clustering via joint between-within distances: Extending ward's minimum variance method, J. Classification 22 (2) (2005) 151–184.
[92] Tessa E.S. Charlesworth, Mahzarin R. Banaji, Gender in science, technology, engineering, and mathematics: Issues, causes, solutions, J. Neurosci. 39 (37) (2019) 7228–7243.
[93] Gianmario Voria, Stefano Lambiase, Maria Concetta Schiavone, Gemma Catolino, Fabio Palomba, From expectation to habit: Why do software practitioners adopt fairness toolkits? in: IEEE/ACM 47th International Conference on Software Engineering (ICSE25), 2024, arXiv preprint arXiv:2412.13846.
[94] Gianmario Voria, Rebecca Di Matteo, Giammaria Giordano, Gemma Catolino, Fabio Palomba, Data preparation for fairness-performance trade-offs: A practitioner-friendly alternative?, 2024, arXiv preprint arXiv:2412.15920.
[95] Ronnie de Souza Santos, Felipe Fronchetti, Savio Freire, Rodrigo Spinola, Software fairness debt, 2024, arXiv preprint arXiv:2405.02490.
[96] T. Nemoto, D. Beglar, Likert-scale questionnaires, in: JALT 2013 Conference Proceedings, 2014, pp. 1–8.
[97] Katherine J. Hunt, Natalie Shlomo, Julia Addington-Hall, Participant recruitment in sensitive surveys: a comparative trial of 'opt in'versus 'opt out'approaches, BMC Med. Res. Methodol. 13 (1) (2013) 1–8.
[98] James J. Heckman, Selection bias and self-selection, in: Econometrics, Springer, 1990, pp. 201–224.
[99] Joseph W Sakshaug, Alexandra Schmucker, Frauke Kreuter, Mick P Couper, Eleanor Singer, Evaluating active (opt-in) and passive (opt-out) consent bias in the transfer of federal contact data to a third-party survey agency, J. Surv. Stat. Methodol. 4 (3) (2016) 382–416.
[100] Felipe Ebert, Alexander Serebrenik, Christoph Treude, Nicole Novielli, Fernando Castor, On recruiting experienced GitHub contributors for interviews and surveys on prolific, in: International Workshop on Recruiting Participants for Empirical Software Engineering, 2022.