From Expectation to Habit: Why Do Software Practitioners Adopt Fairness Toolkits?

Gianmario Voria
University of Salerno
Salerno, Italy
gvoria@unisa.it

Stefano Lambiase
University of Salerno
Salerno, Italy
slambiase@unisa.it

Maria Concetta Schiavone
University of Salerno
Salerno, Italy
m.schiavone29@studenti.unisa.it

Gemma Catolino
University of Salerno
Salerno, Italy
gcatolino@unisa.it

Fabio Palomba
University of Salerno
Salerno, Italy
fpalomba@unisa.it

***Abstract*—As the adoption of machine learning (ML) systems continues to grow across industries, concerns about fairness and bias in these systems have taken center stage. Fairness toolkits—designed to mitigate bias in ML models—serve as critical tools for addressing these ethical concerns. However, their adoption in the context of software development remains underexplored, especially regarding the cognitive and behavioral factors driving their usage. As a deeper understanding of these factors could be pivotal in refining tool designs and promoting broader adoption, this study investigates the factors influencing the adoption of fairness toolkits from an individual perspective. Guided by the Unified Theory of Acceptance and Use of Technology (UTAUT2), we examined the factors shaping the intention to adopt and actual use of fairness toolkits. Specifically, we employed Partial Least Squares Structural Equation Modeling (PLS-SEM) to analyze data from a survey study involving practitioners in the software industry. Our findings reveal that performance expectancy and habit are the primary drivers of fairness toolkit adoption. These insights suggest that by emphasizing the effectiveness of these tools in mitigating bias and fostering habitual use, organizations can encourage wider adoption. Practical recommendations include improving toolkit usability, integrating bias mitigation processes into routine development workflows, and providing ongoing support to ensure professionals see clear benefits from regular use.**

***Index Terms*—Machine Learning Fairness; UTAUT; Technology Adoption; Empirical Software Engineering.**

I. Introduction

Machine Learning (ML) has become pervasive, with its adoption accelerating across a wide array of industries and everyday applications [1]. ML-enabled systems—software systems powered by AI or ML algorithms [2]—are revolutionizing sectors such as healthcare and entertainment by improving efficiency, optimizing decision-making processes, and driving innovative solutions [3]–[6].

As ML continues to spread, it has also prompted significant ethical concerns, particularly around *fairness* [7], which refers to the principle that models should make impartial decisions, avoiding bias or discrimination against certain groups. Unfairness occurs when models inherit biases present in training data [8], [9], resulting in decisions that undermine trust and pose ethical as well as legal challenges [10]. This is proven by several known ethical incidents caused by ML applications, e.g., Facebook vision model that put the "primate" label to black men or Amazon assigning lower sales ranking to books containing LGBTQIA+ themes, highlighting the urgent need for fair ML software [11]–[14].

Acknowledging the critical importance of fairness, the software engineering (SE) research community—particularly in the domain of software engineering for artificial intelligence (SE4AI)—has made substantial strides in developing bias mitigation techniques [15], recognizing fairness as a crucial non-functional requirement. These approaches can generally be grouped into three main categories: *pre-processing*, *in-processing*, and *post-processing* techniques. In this regard, the research community and organizations have developed instruments to make these solutions available for *software practitioners*, e.g., AIF360 [16] or FAIRLEARN [17]. These tools referred to as *fairness toolkits*, comprise ready-to-use metrics to measure fairness or bias mitigation techniques [18].

While fairness toolkits have proven effective in mitigating bias [18], [19], there is still a significant gap in understanding their actual adoption. Specifically, it remains unclear what *decision-making heuristics* lead practitioners to consider using fairness toolkits in their workflow. We argue that this is an important limitation for two reasons. First, studying the adoption of fairness toolkits may uncover the main drivers that need to be considered or encouraged to further increase the uptake of these tools, as suggested by previous research investigating technology acceptance [20], [21]. Second, understanding the considerations that lead to their adoption can offer additional insights into how existing fairness toolkits can be refined and better integrated into practitioners' workflows [22]. This may inform recommendations for designing the next generation of fairness toolkits. In summary, a deeper understanding of these heuristics could provide valuable insights for both researchers and toolkits vendors.

Recognizing the aforementioned opportunities, our goal is to offer a complementary perspective by investigating the key factors influencing practitioners' willingness to adopt fairness toolkits. Therefore, this research seeks to address this gap, starting by defining the following guiding research question:

> **Research Question.** *What factors influence software practitioners in the adoption of fairness toolkits?*

To address the research question, we conducted a quantitative study grounded in the Unified Theory of Acceptance

and Use of Technology (UTAUT2) [23]. We surveyed expert practitioners and employed Partial Least Squares Structural Equation Modeling (PLS-SEM) for data analysis [24]. Our results show that software practitioners' intention to adopt fairness toolkits is mainly driven by their expectancy of the performance of these instruments, i.e., the extent to which they are able to mitigate bias. Moreover, habit emerged as a driver for both the intention to use and the actual adoption of fairness toolkits by practitioners.

II. Background and Related Works

In the following subsections, we summarize the most relevant literature regarding fairness.

Machine Learning Fairness. Fairness in decision-making refers to the absence of bias or favoritism based on inherent or acquired characteristics [7], [9], [25]. Various metrics and strategies evaluate fairness in ML, focusing on data similarities, decision probabilities, and cause-effect relationships [26]. Majumder et al. [27] categorized fairness metrics into seven groups, although not all nuances are captured.

Bias in ML systems can be mitigated using pre-, in-, and post-processing techniques. Pre-processing targets bias in training data, with Sharma et al. [28] and Calmon et al. [29] using probabilistic methods, and Chakraborty et al. [30] introducing FAIR-SMOTE. In-processing adjusts the learning algorithm itself; Zhang et al. [31] used adversarial methods, and Kamishima et al. [32] applied regularization. Reweighting methods, like those from Kamiran and Calders [33] and Chakraborty et al. [34], adjust instance weights. Post-processing techniques refine model outputs after training, with Galhotra et al. [35] introducing THEMIS and Udeshi et al. [36] developing AEQUITAS.

Software Engineering for ML Fairness. Fairness in ML has gained traction in the SE community, with various studies addressing it from multiple perspectives [7], [9], [15], [25], [37]. Ferrara et al. [38] stressed the importance of context-aware fairness requirements. Additionally, Ferrara et al. [39] advocated for fairness integration across the development lifecycle. Discrimination often stems from biased training datasets [40]. Zhang and Harman [41] argued that increasing dataset features does not inherently reduce discrimination, while Chakraborty et al. [30] emphasized the role of feature selection. Sesari et al. [42] highlighted the need to examine fairness across the entire dataset, and Voria et al. [43], [44] cataloged fairness-aware practices surveying domain experts.

Fairness Toolkits in Practice. Various open-source fairness toolkits help researchers and developers create fairer ML models [45]. AIF360 [16], developed by IBM, offers a comprehensive suite of fairness metrics and bias mitigation techniques. FAIRLEARN [17] is a Python-based library focused on fairness assessment and mitigation using in-processing techniques. Google's WHAT-IF TOOL [46] emphasizes fairness and explainability through interactive analysis, while SCIKIT-FAIRNESS [47] extends scikit-learn with bias analysis tools.

Lee and Singh [19] examined the misalignment between current open-source fairness toolkits and practitioners' needs through focus groups, interviews, and surveys, identifying gaps that necessitate improved support for implementing fairness. Similarly, Holstein et al. [48] documented challenges in developing fair ML systems within commercial teams based on interviews and surveys. Deng et al. [18] conducted an empirical investigation into industry practitioners' engagement with fairness toolkits, identifying usability and effectiveness improvements through think-aloud interviews and surveys. Abstracting from the technical perspective, Rakova et al. [22] explored fairness issues organizationally, developing a framework to analyze how organizational culture and structure impact responsible software initiatives. They identified challenges and enablers through interviews, mapping current structures to ideal future processes.

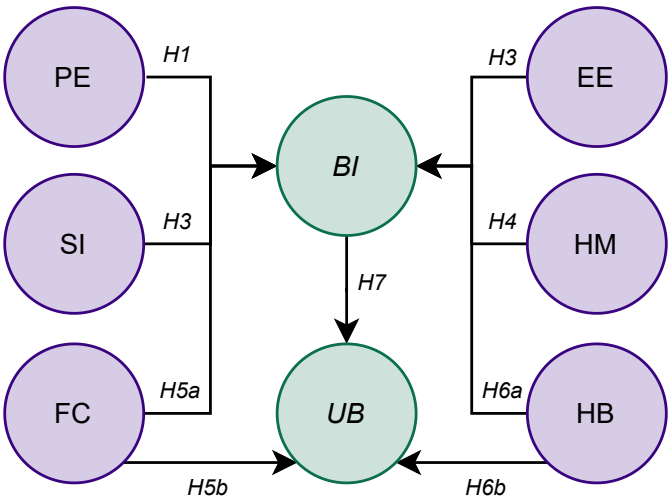

Fig. 1. Overview of the UTAUT2 theoretical model with our hypothesis.

Our Contribution.

The objective of our work aligns with existing studies, specifically focusing on the practical use and effectiveness of fairness toolkits. However, it distinguishes itself by incorporating the UTAUT (Unified Theory of Acceptance and Use of Technology) model as a framework to systematically understand how these tools are utilized and perceived by users. This integration allows for an examination of user adoption and acceptance of technology, along with the identification of factors that influence the effectiveness and usability of fairness toolkits, thus filling a crucial gap in the current literature.

III. Hypothesis and Theory Development

This study seeks to identify the key individual factors that influence the adoption of fairness toolkits in software development and engineering. Given the lack of studies that specifically investigate the factors influencing the adoption of fairness toolkits, we have chosen to base our approach on the UNIFIED THEORY OF ACCEPTANCE AND USE OF

TECHNOLOGY (UTAUT2) [23] model. This theory is widely regarded as one of the most comprehensive frameworks for examining technology adoption across diverse contexts [20], [21], [23], [49], enabling us to minimize bias and build on established knowledge.

A. The Unified Theory of Acceptance and Use of Technology

One of the theoretical models developed to predict technology adoption and use is UTAUT [50]. This model posits that the actual usage **(UB)** of technology is determined by behavioral intention **(BI)**. The perceived likelihood of adopting the technology is influenced by the direct effects of four key constructs: the belief that using the system will enhance job performance (Performance Expectancy, **PE**), the perceived ease of using the system (Effort Expectancy, **EE**), the perception that organizational and technical infrastructures are in place to support system use (Facilitating Conditions, **FC**), and the perception that important others believe the system should be used (Social Influence, **SI**). Individual-related factors, such as age, gender, and experience, are typically considered to moderate or diminish the relationship between technology use and behavioral intention [51].

Despite the widespread acceptance of UTAUT, Venkatesh et al. [23] later introduced UTAUT2, an updated version of the original model that includes three additional constructs, emphasizing the user as a customer and stakeholder rather than merely a technology adopter [23]. This extension was designed to provide greater precision in explaining user behavior. UTAUT2 provides the following additional constructs: the degree of pleasure or enjoyment derived from using the technology (Hedonic Motivation, **HM**), the cognitive trade-off between the perceived benefits of the technology and its monetary cost (Price Value, **PV**), and the extent to which individuals tend to perform behaviors automatically through learning (Habit, **HB**).

Motivation and Choices. Given our objective to investigate the adoption of fairness toolkits by software practitioners, the UTAUT2 model was a natural selection. In comparison to its predecessors, the Technology Acceptance Model (TAM) [52] and UTAUT [50], the new model encompasses a broader range of individual-level factors that capture various dimensions of technology adoption [23]. Moreover, the constructs of social influence and facilitating conditions allow us to consider environmental factors that may affect the adoption process [23]. Importantly, UTAUT2 provides us with established and validated measurement instruments, enabling us to contextualize our findings within a substantial body of literature utilizing the same theoretical framework.

The standard UTAUT2 model includes three moderating variables: age, gender, and experience. However, in the interest of model parsimony, previous research [20], [49] that adopted the UTAUT framework in software engineering research chose to exclude these three. To ensure reliability and robustness, we still conducted a preliminary analysis using the moderating variables. The results, available in our online appendix [53], revealed that these three were not significant. Hence, according to previous research, we made the same decision to exclude the moderators from our final analysis.

Finally, the construct Price Value (PV) [23] has been excluded. It evaluates the perception of the relationship between the price paid and the benefits gained from using the technology. However, since fairness toolkits are mostly open-source and thus freely accessible without any direct economic cost to the user, this construct was deemed insignificant.

B. Hypotheteses Development

In the following, we present the hypothesis we developed for the constructs of the UTAUT2 model to understand software practitioners' intention to use and actual usage of fairness toolkits. Figure 1 summarizes the model built upon the hypotheses described in this section.

One of the constructs of the model is Performance Expectancy, which refers to the degree to which an individual believes that adopting a particular technology will enhance their job performance [50]. In essence, this construct suggests that practitioners are more likely to utilize fairness toolkits if they perceive these tools as beneficial for completing their routine software development tasks, which in this case may be influenced by how much they are expected to produce fair software for their organization [22]. Positive outcomes associated with fairness toolkits—such as improved efficiency in identifying biases, enhanced accuracy in decision-making, and strengthened capabilities for addressing ethical dilemmas—can significantly influence practitioners' willingness to integrate these tools into their workflows [54]. Given the potential of fairness toolkits to streamline development processes and provide substantial performance benefits, it is reasonable to propose that performance expectancy plays a role in practitioners' intentions to adopt these technologies [54], [55]. Therefore, we hypothesize that ***(H1: PE→BI)*** *Performance Expectancy (PE) positively influences the intention to adopt (BI) fairness toolkits by software practitioners.*

The degree of ease with which a particular technology can be utilized is referred to as Effort Expectancy [50]. Individuals are more likely to adopt new technologies when they find them easy to understand and use [56]. Stakeholders may decide whether to incorporate fairness toolkits into their development processes based on how straightforward it is to integrate and utilize these tools [18]. We anticipate that effort expectancy will positively influence the intention to adopt fairness toolkits for software development, given the significance of ease of use and reduced cognitive load in technology acceptance. Therefore, we hypothesize that ***(H2: EE→BI)*** *Effort Expectancy (EE) positively influences the intention to adopt (BI) fairness toolkits by software practitioners.*

Social Influence refers to the degree to which an individual perceives that important others believe they should use a new system [50]. According to this concept, individuals are more likely to adopt new technologies if they feel that their colleagues, supervisors, or social norms advocate for their use. Stakeholders may receive approval and encouragement from peers, mentors, or industry leaders, which can significantly

impact their decision to adopt fairness toolkits [54], given the critical impact that fairness may have on society. Therefore, we hypothesize that ***(H3: SI→BI)*** *Social Influence (SI) positively influences the intention to adopt (BI) fairness toolkits by software practitioners.*

Hedonic Motivation refers to the pleasure or enjoyment derived from using fairness toolkits for development tasks [23]. This construct significantly influences user acceptance of technology. It can be observed that the more enjoyable an activity is, the more positive the practitioner's attitude toward it becomes [57]. Therefore, we hypothesize ***(H4: HM→BI)*** *Hedonic Motivation (HM) positively influences the intention to adopt (BI) fairness toolkits by software practitioners.*

The belief of an organizational and technical infrastructure that supports the use of a particular technology is known as Facilitating Conditions [50]. Practitioners are more likely to adopt fairness toolkits if they have access to the necessary tools, as well as support and training. They are also more inclined to actually utilize them if they are confident that their organization provides the resources to integrate and use them effectively. Therefore, we hypothesize that *Facilitating Conditions (FC)* ***(H5a: FC→BI)*** *positively influences the intention to adopt (BI) fairness toolkits by software practitioners and* ***(H5b: FC→UB)*** *positively influences the actual use behavior (UB) regarding fairness toolkits by practitioners.*

Habit refers to the extent to which individuals tend to act automatically as a result of learning [23]. Practitioners are more likely to adopt fairness toolkits if they are familiar with certain tools and use them regularly. Furthermore, habit influences actual usage behavior, as practitioners who consistently integrate new tools are more inclined to use fairness toolkits consistently in their development activities. Therefore, we hypothesize that *Habit (HB)* ***(H6a: HB→BI)*** *positively influences the intention to adopt (BI) fairness toolkits by software practitioners and* ***(H6b: HB→UB)*** *positively influences the actual use behavior (UB) regarding fairness toolkits by software practitioners.*

Lastly, according to psychological models, individual behavior can be predicted and conditioned by personal intentions. Hence, UTAUT2 posits that behavioral intention significantly impacts actual technology usage [50]. Based on this premise, we hypothesize that ***(H7: BI→UB)*** *Behavioral Intention (BI) to use fairness toolkit positively influences the actual use behavior (UB) of software practitioners.*

We combined the hypotheses outlined above with those from the UTAUT2 framework to explore software practitioners' adoption of fairness toolkits, as shown in Figure 1.

IV. Research Design

To evaluate the above-mentioned hypotheses and explore software practitioners' intentions to adopt fairness toolkits, we conducted a survey study.

Figure 2 presents an overview of our research methodology, which we elaborate upon in this section. We initiated our process by carefully defining participant selection criteria for our survey and calculating the sample size using G*Power [58], taking into account the complexities of our theoretical model.

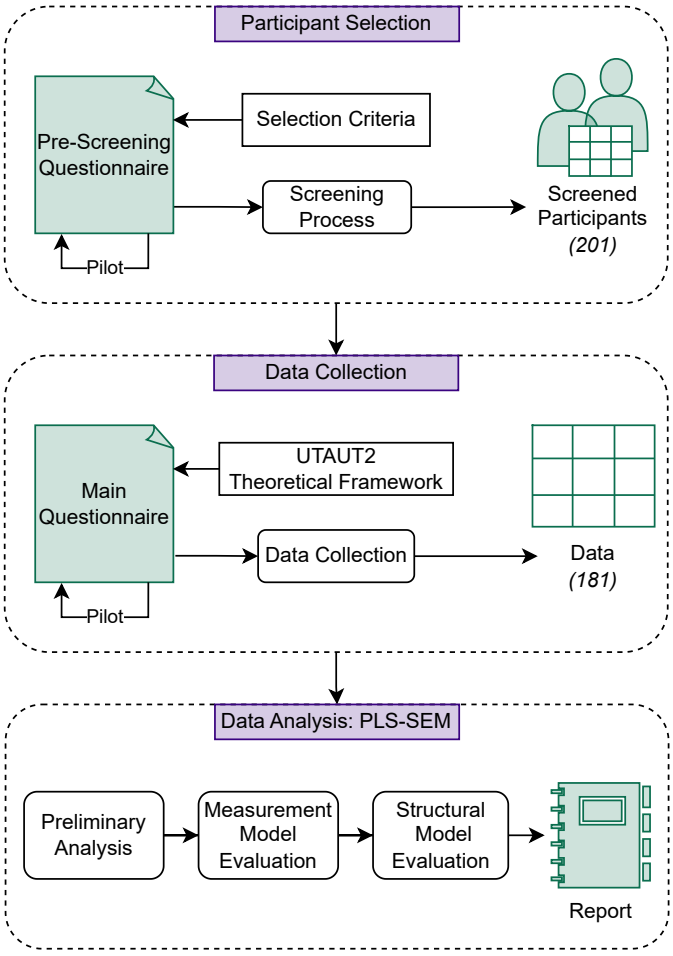

Fig. 2. Research Method.

Subsequently, we developed questionnaires grounded in validated instruments from the literature [23] to measure the model's constructs. Throughout the questionnaire development process, we adhered strictly to established guidelines, incorporating iterative pilot testing to ensure the highest standards of validity and reliability [59], [60].

Participants first completed a carefully designed screening questionnaire, followed by a second survey intended to assess the UTAUT2 constructs. We then rigorously analyzed the collected data using *Partial Least Squares-Structural Equation Modeling (PLS-SEM)* [24] to address our research questions. PLS-SEM is a versatile statistical method that integrates factor analysis and multiple regression to analyze complex relationships between constructs [24], [61]. It is especially useful for predictive analysis, theory development, and models with multiple independent and dependent variables, particularly when dealing with small sample sizes or non-normal data distributions.

A. Participant Selection and Demographics

To gather data for this study, we implemented a survey utilizing a cluster sampling strategy via Prolific, a reputable

academic data collection platform.[1] Prolific's advanced filtering capabilities allowed us to precisely define specific criteria for selecting potential participants that aligned with our research requirements.

We formulated an ideal participant profile based on our study's objectives, focusing on individuals demonstrating proficiency in computer programming and actively engaged in the information technology industry. Leveraging Prolific's filters, we targeted participants meeting these exacting criteria. The survey, titled *["Fairness in Software Development] Pre-Screening — Fairness Toolkit Adoption"*, clearly stated its goals in the description, ensuring that only qualified respondents, including those not currently employed as developers but meeting the pre-screening standards, participated.

To further enhance the reliability and accuracy of our data, we implemented two additional filters. First, we required that participants be fluent in English, as the survey was conducted exclusively in this language. Second, we required participants to have a 100% approval rate from previous surveys. On Prolific, researchers can reject responses based on specific criteria, and a participant's approval rate reflects the percentage of submissions that have been accepted, serving as a reliable proxy for their consistency in providing high-quality data. In addition to these filters, we incorporated questions developed by Danilova et al. [62] to assess participants' programming knowledge, ensuring they possessed the required expertise. Lastly, we prioritized recruiting individuals with a high degree of familiarity with fairness and its related toolkits in a professional context; this was done using custom screening directly in the questionnaire.

Based on (most of) the criteria mentioned above, Prolific identified a pool of 3046 potential participants. We determined the minimum sample size by conducting *a priori* power analysis using G*Power [58]. With an effect size of 15%, a significance level of 5%, and a power of 95%, we calculated the smallest required sample size for seven predictors to be 153 participants. To mitigate potential dropout between the pre-screening and the main survey, we initially collected 201 responses and subsequently received 181 responses for the main survey. The time gap between the two surveys facilitated thorough participant screening. The larger initial pool ensured us a sufficient sample size for analysis, anticipating that some participants might not proceed to the main survey.

Our final sample comprised 181 participants, exhibiting a demographic distribution of 80% men, 19% women, and 1% non-binary individuals. The participants came from 15 diverse countries, with the largest contingents originating from the United States (39%) and the United Kingdom (24%). Other nations represented included Australia, Belgium, Canada, Finland, France, Germany, Ireland, Italy, the Netherlands, Portugal, Singapore, Spain, and Sweden, ensuring a rich diversity of perspectives in our study. Our survey encompassed a diverse array of work positions among the 181 respondents. The most prevalent roles included Software Developer/Programmer (25%), Project Manager (15%), Data Analyst/Data Engineer/Data Scientist (14%), and Software Engineer (13%). The respondents also reflected a broad spectrum of ages and experience levels. The majority (52%) fell within the 30 to 44 years age bracket, followed by 25% in the 18 to 29 range, and 20% between 45 and 59 years old. A mere 3% were older than 69 years. Regarding experience in the software industry, 30% had over 10 years of experience, 28% had 1-2 years, and 24% had 4-6 years. A smaller sample reported 7-9 years of experience (16%), while only 1% had less than 1 year of experience in industry.

[1]Prolific (www.prolific.com) [October 2024]

B. Data Collection

To facilitate data collection, we developed two questionnaires: the first, designated as the "Pre-Screening questionnaire," aimed to identify ideal participants—by mean of custom screening—from those already filtered via the Prolific platform. The second, named the "Main questionnaire," was designed to measure UTAUT2 constructs in the participants selected from the pre-screening survey. Both surveys were developed adhering to the established guidelines by Kitchenham and Pfleeger [59] and Andrews et al. [60], which are highly regarded in software engineering research. Additionally, we followed the SIGSOFT Empirical Standard for Questionnaire Surveys [63]. The questionnaires were fully anonymized, featuring an introductory description that provided key details to aid participants in comprehending the tasks. We incorporated a closing question for feedback and attention check questions to ensure participant reliability.

Before administering the surveys, we conducted iterative pilot tests with dual objectives: (1) assessing quality and clarity and (2) estimating completion time. Initially, we orchestrated three pilots involving 10 researchers from our network. After each round, we refined the surveys to address feedback and address any typographical errors. Subsequently, we conducted separate pilots for each questionnaire using Prolific: five participants completed the pre-screening questionnaire, and another five completed the main one. The pilots for both questionnaires happened between 30 August and 01 September 2024.

The pre-screening questionnaire gathered demographic information, assessed participant reliability, and evaluated programming skills and experience. Designed to take 6 minutes to complete, we successfully collected 200 responses within two days starting from 30 August 2024. The main questionnaire, which measured the UTAUT2 constructs, was estimated to require 5 minutes for completion, and it took six days to collect 181 responses from 01 September 2024.

Ethical Considerations. We thoughtfully designed and executed our work, giving careful consideration to participants' privacy and addressing potential ethical concerns inherent in survey studies [64]. Our survey design ensured complete anonymity of all responses; consequently, we refrained from collecting participants' names or email addresses. We avoided soliciting any sensitive business information and explicitly guaranteed that the collected data would be utilized solely to address our research objectives. All participants were over 18

years old, provided informed consent prior to participation, and were allowed to withdraw at any time. Moreover, we transparently informed participants that their responses would eventually be published and permanently stored in the online appendix of this paper [53].

Nevertheless, we acknowledge that gathering insights on critical aspects—such as fairness—from potential employees of organizations that could produce discriminatory ML-based products may still present moral concerns. However, we recognize that industry practitioners have been involved in evaluating fairness and ethics in previous work [18], [22]. Furthermore, given that the scope of the survey was clearly presented in the introduction, we are confident that all participants who responded were genuinely motivated to pursue the cause of providing non-discriminatory solutions.

C. Data Analysis

As previously explained, data collection was conducted through a survey study. All constructs in the theoretical model described in Section III were measured at the individual level using items validated in the literature, ensuring the reliability of our measurement process. We detail the items used and the data analysis process in the following sections, with a comprehensive overview of all items and references provided in our online appendix [53].

1) Data Gathering Instruments: Questionnaire items assessing the UTAUT2 constructs were adapted from the original authors [23]. The dependent variable, *Use Behavior* (UB), was measured using a single-item frequency scale, while the seven predictors were evaluated on a 7-point Likert scale. These predictors encompassed *Performance Expectancy* (PE, 5 items), *Effort Expectancy* (EE, 6 items), *Social Influence* (SI, 5 items), *Hedonic Motivation* (HM, 3 items), *Facilitating Conditions* (FC, 4 items), *Habit* (HB, 4 items), and *Price Value* (PV, 3 items), as well as *Behavioral Intention* (BI, 3 items). We also asked whether the use of fairness toolkits was mandated by the participants' companies, recognizing this as a potential influencer of use behavior. Additionally, we collected demographic data such as age, gender, role, and years of experience in the software industry to contextualize our sample relative to other surveys.

2) Analysis Process: We initiated our analysis by conducting a thorough preliminary examination of the data to ensure its quality. While PLS-SEM offers considerable flexibility, we nonetheless checked for missing data, unusual response patterns, outliers, and data distribution issues. Upon validating the dataset, we imported it into SmartPLS, a tool designed specifically for PLS-SEM analysis [65]. Given the intricate nature of the PLS-SEM process, we direct readers to Hair et al. [24] and Russo and Stol [61] for more detailed explanations.

We began by developing the measurement model (or outer model), which links each theoretical construct to its associated indicators. Each construct in our theoretical framework, i.e., the UTAUT2 values, was treated as a latent variable, representing an unobservable concept. Indicators derived from participants' responses were then carefully assigned to their respective constructs. Subsequently, we constructed the structural model (or inner model), which delineates the relationships between these constructs, guided by our hypotheses.

After constructing both the measurement and structural models within SmartPLS and executing the PLS-SEM algorithm, we evaluated both. Given that all our indicators exhibited a reflective relationship with the constructs, we assessed the criteria indicated by Hair et al. [24].

To support replicability and maintain transparency, all materials utilized during the analysis are provided in the comprehensive replication package accompanying this paper [53].

V. Analysis of the Results

The following section presents the results from the PLS-SEM analysis. This analysis aims to uncover the causal relationships and underlying patterns within the data, offering a detailed evaluation of the hypothesized model. Through this approach, we gain insights into the interactions between the constructs and validate the theoretical framework proposed.

Before proceeding with the main PLS-SEM analysis, we conducted a preliminary data examination. Notably, there were only a few instances of missing values, likely due to the high quality of the questionnaire and the reliability of our sample, ensured by the approval rate filter. These missing values did not pose any significant issues, as SmartPLS is equipped to manage them automatically. Additionally, we reviewed the data for suspicious response patterns and found none. It is also important to highlight that all participants successfully passed the attention check questions.

A. Measurement Model Evaluation

As a first step in the evaluation of the theoretical model, it is paramount to evaluate the reliability of the constructs of the model [24], [61]. Consequently, we analyze the indicator reliability, internal consistency reliability, convergent validity, and discriminant validity. This section presents the obtained results for each of the steps mentioned above.

Indicator Reliability. As outlined by Hair et al. [24], the initial step in assessing the measurement model is to evaluate the reliability of the indicators, focusing on their *outer loadings*. High outer loadings indicate that the indicators capture a substantial amount of commonality with the construct.

A commonly accepted guideline is to retain indicators with outer loadings above 0.708, while indicators with loadings below 0.40 are generally removed. For those with values between 0.40 and 0.70, removal is considered if it enhances internal consistency reliability or convergent validity.

For brevity, the outer loadings for all indicators are reported in the online appendix [53]. Two indicators, EE4 and HB2, had outer loadings below 0.70, but since no indicator had a loading lower than 0.40, all were retained for further analysis.

Internal Consistency Reliability. The second step involved evaluating *internal consistency reliability* to confirm that the indicators reliably measure their respective constructs. For this assessment, we used three key measures: *Cronbach's alpha*, *composite reliability* (ρ_c), and the *reliability coefficient* (ρ_A).

TABLE I
INTERNAL CONSISTENCY RELIABILITY AND AVE VALUES OF THE UTAUT2 CONSTRUCTS.

Constructs	Cronbach's alpha	ρ_A	ρ_c	AVE
BI	0.888	0.890	0.931	0.817
EE	0.868	0.887	0.899	0.598
FC	0.817	0.829	0.879	0.647
HB	0.841	0.866	0.894	0.681
HM	0.920	0.937	0.950	0.863
PE	0.915	0.921	0.936	0.746
SI	0.866	0.903	0.900	0.644

The results, presented in Table I, show that all values exceeded the recommended threshold of 0.60, as suggested by Hair et al. [24], allowing us to pass this step confidently.

Convergent Validity. Convergent validity refers to the degree to which a measure correlates positively with other measures of the same construct [24]. Since all constructs in our model utilize reflective indicators, we anticipated that the indicators would converge and share a substantial proportion of variance. The most common metric for assessing this is the *average variance extracted* (AVE). An AVE value of 0.50 or higher is considered acceptable, indicating that the construct captures more than half of the variance of its associated indicators. As reported in Table I, all constructs in our model have AVE values exceeding the 0.50 threshold, confirming strong convergent validity.

Discriminant Validity. The final test focused on assessing the discriminant validity, which evaluates the degree to which a construct is truly distinct from others. Henseler et al. [66] introduced the *heterotrait-monotrait ratio* (HTMT) as a reliable criterion for this purpose. HTMT values are computed using the PLS-SEM algorithm, and typically, a value above 0.90 indicates insufficient discriminant validity, while values below 0.85 suggest it is adequate. Additionally, using a *bootstrapping* procedure can further verify whether the HTMT values significantly differ from the threshold. Bootstrapping is a nonparametric technique used to test the significance of various PLS-SEM outcomes, such as path coefficients, Cronbach's alpha, and HTMT values.

Due to space constraints, the HTMT values derived from our PLS-SEM analysis are included in the online appendix of this paper [53]. The results demonstrated that all values fell below the 0.85 threshold. We performed a bootstrapping procedure in SmartPLS with *10000* subsamples, using a one-tailed test at a 0.05 significance level, which confirmed that all HTMT values were below the thresholds. These findings indicate that each construct in our model represents a distinct concept, allowing us to proceed with the evaluation of the structural model.

B. Structural Model Evaluation

After evaluating the measurement model, the next step is to assess the structural model.

Collinearity Analysis. The initial step in evaluating the structural model involves examining collinearity between exogenous and endogenous variables, which is essential for accurate path estimation. To detect multicollinearity, we employed the *Variance Inflation Factor* (VIF), a standard metric used in multiple regression analysis. Ideally, a VIF value under 3 indicates no collinearity, while values below 5 are also acceptable. In our analysis, the majority of VIF values were below 3, with the highest being 2.47. Only two paths (PE $\rightarrow$ BI and EE $\rightarrow$ BI) slightly exceeded the ideal threshold. Based on these findings, we concluded that multicollinearity does not present a significant concern in our model.

Significance and Relevance of the Relationships. In the second phase of our analysis, we focused on evaluating the significance and relevance of the relationships within the structural model. To test for significance, we employed the bootstrapping method, using 10,000 sub-samples, as recommended by Hair et al. [24]. We analyzed T-values, p-values, and bootstrap confidence intervals. The results, summarized in Table II, indicate that Habit significantly influences both Use Behavior and Behavioral Intention. Furthermore, Behavioral Intention is strongly associated with Use Behavior, while Performance Expectancy demonstrates a significant connection to Behavioral Intention. To assess the relevance of these significant relationships, we examined the standardized path coefficients, which are also detailed in Table II. Performance Expectancy emerged as the most influential factor affecting the intention to utilize fairness toolkits, closely followed by Habit. In terms of actual usage behavior among software practitioners, Habit was identified as the most significant factor, with Behavioral Intention ranking second. In addition, further analysis revealed that Performance Expectancy also has an indirect relationship with the use behavior.

Explanatory Power. In the third phase of our analysis, we aimed to evaluate the model's explanatory capability, specifically how well it fits the data by quantifying the strength of the relationships within the model, as described by Hair et al. [24] and Russo et al. [61]. This is typically assessed using the coefficient of determination (R^2), which ranges from 0 to 1; higher values indicate stronger explanatory power. Although no universal standards exist for R^2, values as low as 0.10 may be considered acceptable in certain contexts, with 0.19 often regarded as a more appropriate benchmark [24], [67], [68].

In our analysis, we found R^2 values of 0.630 for Behavioral Intention and 0.407 for Use Behavior. This indicates that our model successfully explains 63% of the variance in the intention to use large language models (LLMs) and 40% of the variance in actual usage. Furthermore, since R^2 values are below 0.90, we can confidently exclude the overfitting concern.

After evaluating the coefficient of determination, we further quantified the strength of the relationships using the F^2 effect size. This metric assesses the potential change in R^2 if a specific construct were omitted from the model, offering insights into the individual contributions of each construct to the dependent variables.

TABLE II
SIGNIFICANCE AND RELEVANCE OF THE UTAUT2 CONSTRUCTS. SIGNIFICANT PATHS ARE HIGHLIGHTED IN BOLD AND UNDERLINED.

Hypotheses	Path Coefficients	Bootstrap Mean	St. Dev	T statistics	P values	Significance
BI → UB	0.161	0.159	0.071	2.258	0.024	*
EE → BI	0.056	0.057	0.090	0.627	0.531	
FC → BI	0.090	0.098	0.077	1.169	0.242	
FC → UB	0.091	0.092	0.054	1.688	0.091	
HB → BI	0.323	0.319	0.091	3.531	0.000	**
HB → UB	0.543	0.455	0.075	6.014	0.000	**
HM → BI	-0.067	-0.064	0.078	0.852	0.94	
PE → BI	0.465	0.465	0.086	5.404	0.000	**
SI → BI	0.011	0.010	0.065	0.174	0.862	

$**: p<0.001$; $*: p<0.05$

Focusing on the intention to use fairness toolkits, we found that Performance Expectancy had the largest effect size (0.192), followed by Habit (0.096). When considering actual usage, Habit was the most influential factor (0.150), with Behavioral Intention showing a smaller effect size (0.021). These results highlight the significant influence of Performance Expectancy on the intention to adopt fairness toolkits and underscore the critical role of habitual usage in predicting actual usage.

Predictive Power. To evaluate the model's practical utility for managerial decision-making, we assessed whether the results derived from our PLS-SEM algorithm are generalizable beyond the specific dataset used in the estimation. This was done using the $PLS_{predict}$ procedure [69], which divides the dataset into training and holdout samples. The key metrics in this analysis were Stone-Geisser's Q^2 *statistic*, along with the *mean absolute error* (MAE) and the *root mean square error* (RMSE). These values were compared to a benchmark, with Shmueli et al. [69], [70] recommending a linear regression model (LM) as the benchmark for comparison. A positive Q^2 value indicates that the model's prediction error is lower than that of the benchmark, while smaller MAE and RMSE values suggest that the model has superior predictive accuracy.

The results, detailed in our online appendix [53] due to space constraints, show that all variables outperform the benchmark, indicating that the PLS-SEM model demonstrates strong predictive capabilities.

Summary of the Results.

The evaluation of the measurement and structural model confirmed the robustness of the data collection process and allowed us to answer our research question by identifying three key factors driving the adoption of fairness toolkits: Performance Expectancy, Habit, and Behavioral Intention.

VI. DISCUSSION AND IMPLICATIONS

This study aimed to explore individual factors influencing software practitioners' intention to adopt, as well as their actual adoption, of fairness toolkits using the Unified Theory of Acceptance and Use of Technology (UTAUT) framework. Our findings indicate that three key constructs from the UTAUT2 [23] model—Performance Expectancy, Habit, and Behavioral Intention—exert a statistically significant influence on the dependent variables. Conversely, the other constructs did not show significant effects on the dependent variables. The remainder of this section will explore and elaborate on all the constructs, offering insights and implications that could be valuable for further research and practice.

A. Discussions

Performance Expectancy, or the perceived utility of fairness toolkits, emerges as the cornerstone of adoption. As fairness becomes an increasingly critical non-functional requirement in modern software engineering [7], [11], practitioners are drawn to tools that effectively mitigate bias in ML systems. Our results reveal that this factor has the most significant influence on the intention to adopt fairness toolkits while also maintaining a strong and significant indirect relationship with the actual use behavior of practitioners. These results align with established technology acceptance models [50] and recent research in the field [54], [55].

The consideration of fairness in daily workflows is largely driven by *Habit*. The study identifies this as the second most influential factor on *behavioral intention* and the primary determinant of *actual use behavior*.[2] This finding underscores the importance of seamless integration and initial exposure in fostering sustained use of fairness toolkits, as habitual use becomes an essential part of practitioners' routines.

The intentional adoption of fairness toolkits is reflected in the strong relationship between *Behavioral Intention* and *actual use behavior*. This connection, consistent with technology

[2]Notably, mediation analysis shows that Habit's influence on use behavior is direct and not mediated by *behavioral intention*.

TABLE III
UTAUT2—SUMMARY OF FINDINGS AND IMPLICATIONS. SIGNIFICANT PATHS ARE HIGHLIGHTED IN BOLD AND UNDERLINED.

Hypothesis	Findings	Implications
H1: PE → BI	This relationship is the strongest of the model regarding the Behavioral Intention, with a path coefficient of 0.465 and an effect size of 0.192. Moreover, this relationship also causes a significant indirect relationship between PE and UB.	Software practitioners' intentions to adopt fairness toolkits are heavily influenced by their expectations of the technology's performance, i.e., how able these instruments are to measure or mitigate bias.
H2: EE → BI	The relationship is not significant.	The perceived effort in learning how to apply fairness toolkits to their jobs does not instigate the intention to adopt them.
H3: SI → BI	The relationship is not significant.	Practitioners' opinion on the use of fairness toolkits alone does not instigate the intention to adopt the tool.
H4: HM → BI	The relationship is not significant.	Practitioners do not intend to adopt fairness toolkits on the basis of the fun and joy their use causes.
H5a: FC → BI	The relationship is not significant.	Organizational support and supporting resources do not significantly influence practitioners' intention to adopt fairness toolkits in their working context.
H5b: FC → UB	The relationship is not significant.	Organizational support and supporting resources do not significantly influence practitioners' actual adoption of fairness toolkits in their working context.
H6a: HB → BI	This relationship is the second most significant regarding Behavioral Intention, with a path coefficient of 0.323 and an effect size of 0.096.	Practitioners that regularly utilize fairness toolkits in their work strengthen their intention to further rely on these tools.
H6b: HB → UB	This relationship is the most significant regarding Use Behavior, with a path coefficient of 0.543 and an effect size of 0.150. A mediation analysis revealed that the effect is direct and not mediated by the relationship with Behavioral Intention.	The habitual use of fairness toolkits can lead to a higher adoption rate.
H7: BI → UB	This relationship is the second most significant regarding Use Behavior: path coefficient of 0.161 and effect size of 0.021.	As expected, the intention to adopt fairness toolkits results in an actual adoption of the technology.

acceptance literature [23], [50], emphasizes that cultivating positive intentions through training, awareness programs, and demonstrating benefits can effectively promote adoption [22].

Notably, the absence of a significant relationship between both *Facilitating Conditions* and *Social Influence* with the intention to adopt and actual use behavior of fairness toolkits suggests that practitioners do not consider organizational support, resource availability, or peer influence as critical factors in their decision to implement these technologies [50], [71]. This finding challenges conventional assumptions about technology adoption in organizational settings, where such factors are typically seen as essential. Instead, the strong influence of *Performance Expectancy* and *Habits* on the dependent variables indicates that software practitioners are primarily motivated by their perception of the toolkits' effectiveness in mitigating bias and their existing work routines. This suggests a self-driven approach to adoption, where practitioners integrate fairness toolkits based on the anticipated positive impact on their tasks rather than relying on organizational facilitation or social endorsement. Practitioners tend to prioritize the tool's utility and performance, further reducing the impact of social factors on their intention to adopt the technology.

Further supporting this interpretation is the non-significant relationship between *Effort Expectancy* and the intent to adopt these toolkits. This finding suggests that software practitioners, given their technical expertise and familiarity with complex tools, prioritize the functionality and performance benefits of fairness toolkits over considerations of ease of use. It appears that the potential for bias mitigation outweighs concerns about the effort required to implement these technologies [18], [19]. Another explanation could lie in the tools themselves, which may offer such a simple interaction that it becomes effortless for professionals to use. Rather than deterring adoption, this complexity seems to be accepted as an inherent aspect of working with cutting-edge AI ethics solutions.

Finally, the study found that *Hedonic Motivation* does not significantly influence the *Behavioral Intention* to adopt fairness toolkits. This finding suggests that the potential enjoyment or pleasure derived from using these tools does not play a substantial role in engineers' adoption decisions [50]. The non-significance of hedonic motivation in this context is particularly noteworthy. It implies that engineers approach fairness toolkits primarily as utilitarian instruments rather than sources of enjoyment or satisfaction. This perspective aligns with the professional nature of software engineering and the ethical implications of fairness in AI systems. Engineers appear to be driven more by the practical outcomes and ethical considerations of using fairness toolkits than by any intrinsic enjoyment derived from the tools themselves.

Our findings may indicate a high level of professional autonomy and ethical responsibility among software engineers working on AI systems. The emphasis on individual assessment and practical utility in adopting fairness toolkits suggests a workforce that is critically engaged with the ethical implications of their work and committed to addressing these issues through technical means.

B. Implications

This study contributes to understanding the reasons behind software practitioners' adoption of fairness toolkits. Our results have actionable implications for organizations, toolkit vendors, and researchers.

Organizations. For organizations that aim to spread the usage of fairness toolkits, these insights suggest that efforts to promote adoption should focus on demonstrating their concrete benefits and effectiveness in addressing bias issues. 👉 *Educational initiatives and awareness campaigns, such as workshops or tutorials*, might be effective if they emphasize the technical merits and the tangible benefits of fairness toolkits rather than relying on social proof, external supports, or attempts to make the tools more enjoyable to use. Moreover, organizations and managers should make an effort to 👉 *integrate fairness toolkits usage into daily workflows to help employees develop a habit*. This can be done by facilitating access and integration of these tools in daily working activities.

Toolkits Vendors. For toolkits vendors, our work may be of inspiration to understand possible design solutions to enhance the adoption of fairness toolkits. To demonstrate toolkits' high performances, vendors should aim to 👉 *provide practical examples and real-world cases* in which their solutions helped mitigate biases and achieve fair ML models, rather than relying on theoretical proofs. In addition, to make practitioners develop a habit of the use of fairness toolkits, vendors should 👉 *facilitate practitioners in integrating such solutions in their daily activities* through efficient APIs or libraries. Despite the efforts vendors can make to promote toolkit adoption, our findings indicate that practitioners perceive these tools as both useful and effective. This suggests that the investment in supporting fair ML development is paying off.

Researchers. Finally, researchers should leverage our findings to perform further investigations on fairness toolkits. On the one hand, 👉 *empirical studies demonstrating these solutions' performances and abilities in mitigating bias* could further increase practitioners' intention to adopt them. On the other hand, exploring 👉 *novel ways to integrate and automate fairness toolkits' integration in existing workflows*, such as CI/CD pipelines, could tempt software practitioners to use them and consequently develop a habit.

VII. Threats To Validity

Our study primarily focused on quantitative analysis supported by statistical methods. In discussing threats to validity, we followed the framework outlined by Wohlin et al. [72].

Regarding the **conclusion Validity**—i.e., threats about the ability to draw accurate conclusions about the relationships between independent and dependent variables [72]—the primary threats in this category stem from the statistical tests used for analysis. To address this, we relied on PLS-SEM, which is known for its robustness in various contexts. We closely followed the procedures outlined by Hair et al. [24] in their detailed work on PLS-SEM methodology. Moreover, we employed SmartPLS, a widely used software cited in over 1,000 peer-reviewed studies [24].

Concerning **internal Validity**—i.e., the risk that external factors may have influenced the dependent variable, leading to inaccurate conclusions [72]—we grounded our study in well-established theories to avoid it. Indeed, we utilized the Unified Theory of Acceptance and Use of Technology, which is specifically suited for investigating phenomena of this nature [23]. We also applied filters to our participant sample to ensure it accurately represented the target population while maintaining sufficient diversity.

Moving on **construct Validity**—that concerns about the accuracy of the measurements and tools used to represent the study variables [72]—all variables were assessed using validated instruments [23]. The questionnaires were designed in line with the latest field guidelines, and we employed strategies such as question randomization and attention-checks to improve the reliability of the results [59], [62], [63].

Last, we tried to address **external Validity**—that is about the generalizability of the findings to a broader population [72] filtering the Prolific population to select participants with characteristics aligned to our study's objectives. Additionally, we gathered sufficient data in line with G*Power recommendations [58]. While the majority of participants were from Europe, which reflects Prolific's user distribution, we acknowledge this limitation. Nonetheless, we believe the results offer valuable insights.

VIII. Conclusion

This study investigated the adoption of fairness toolkits among software practitioners using the Unified Theory of Acceptance and Use of Technology (UTAUT2) framework [23]. We surveyed experts and analyzed the data using Partial Least Squares Structural Equation Modeling (PLS-SEM) [24].

Our findings reveal that *Habit* and *Performance Expectancy* significantly influence the intention to adopt fairness toolkits, aligning with previous research [54], [55]. Moreover, *Habit* emerged as the primary driver for the actual use of these toolkits, alongside practitioners' intention to use them.

These results have important implications. Organizations promoting fairness toolkit adoption should focus on demonstrating their concrete benefits and effectiveness in addressing bias issues. Additionally, our findings suggest that software practitioners primarily approach bias mitigation from a technical perspective, indicating a continued need for research into algorithmic solutions for ML fairness.

Future research should explore the impact of additional factors, such as cultural values, on the adoption of these technologies. We also recommend longitudinal studies to understand how these results may evolve as the technology matures and awareness of AI ethics grows within the software development community.

Data Availability

The data that support the findings of this study are openly available in our online appendix [53].

Acknowledgment

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, Projects PRIN 2022 PNRR "FRINGE: context-aware FaiRness engineerING in complex software systEms" (grant n. P2022553SL, CUP: D53D23017340001). The opinions presented in this article solely belong to the author(s) and do not necessarily reflect those of the European Union or The European Research Executive Agency. The European Union and the granting authority cannot be held accountable for these views.

REFERENCES

[1] J. Zhou and F. Chen, *Human and Machine Learning*. Springer, 2018.

[2] S. Martínez-Fernández, J. Bogner, X. Franch, M. Oriol, J. Siebert, A. Trendowicz, A. M. Vollmer, and S. Wagner, "Software engineering for ai-based systems: A survey," *ACM Transactions on Software Engineering and Methodology*, vol. 31, no. 2, 2022. [Online]. Available: http://dx.doi.org/10.1145/3487043

[3] P. Wang, E. Fan, and P. Wang, "Comparative analysis of image classification algorithms based on traditional machine learning and deep learning," *Pattern recognition letters*, vol. 141, pp. 61–67, 2021.

[4] J. Ni, Y. Chen, Y. Chen, J. Zhu, D. Ali, and W. Cao, "A survey on theories and applications for self-driving cars based on deep learning methods," *Applied Sciences*, vol. 10, no. 8, p. 2749, 2020.

[5] C. C. Miller, "Can an algorithm hire better than a human," *The New York Times*, vol. 25, 2015.

[6] P. Olson, "The algorithm that beats your bank manager," *CNN Money March*, vol. 15, 2011.

[7] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Computing Surveys (CSUR)*, vol. 54, no. 6, pp. 1–35, 2021.

[8] T. P. Pagano, R. B. Loureiro, F. V. Lisboa, R. M. Peixoto, G. A. Guimarães, G. O. Cruz, M. M. Araujo, L. L. Santos, M. A. Cruz, E. L. Oliveira *et al.*, "Bias and unfairness in machine learning models: a systematic review on datasets, tools, fairness metrics, and identification and mitigation methods," *Big data and cognitive computing*, vol. 7, no. 1, p. 15, 2023.

[9] D. Pessach and E. Shmueli, "A review on fairness in machine learning," *ACM Computing Surveys (CSUR)*, vol. 55, no. 3, pp. 1–44, 2022.

[10] S. Miller, "Machine learning, ethics and law," *Australasian Journal of Information Systems*, vol. 23, pp. 1–13, 2019.

[11] Y. Brun and A. Meliou, "Software fairness," in *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.

[12] R. Mac, "Facebook apologizes after a.i. puts 'primates' label on video of black men," *The New York Times*. [Online]. Available: https://www.nytimes.com/2021/09/03/technology/facebook-ai-race-primates.html

[13] B. Johnson and H. Pidd, "'gay writing' falls foul of amazon sales ranking system," *The Guardian*. [Online]. Available: https://www.theguardian.com/culture/2009/apr/13/amazon-gay-writers

[14] M. Wei and Z. Zhou, "Ai ethics issues in real world: Evidence from ai incident database," 2022.

[15] M. Hort, Z. Chen, J. M. Zhang, M. Harman, and F. Sarro, "Bias mitigation for machine learning classifiers: A comprehensive survey," *ACM Journal on Responsible Computing*, vol. 1, no. 2, pp. 1–52, 2024.

[16] R. K. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilović *et al.*, "Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias," *IBM Journal of Research and Development*, vol. 63, no. 4/5, pp. 4–1, 2019.

[17] "Fairlearn," 2019. [Online]. Available: https://fairlearn.github.io/

[18] W. H. Deng, M. Nagireddy, M. S. A. Lee, J. Singh, Z. S. Wu, K. Holstein, and H. Zhu, "Exploring how machine learning practitioners (try to) use fairness toolkits," in *2022 ACM Conference on Fairness, Accountability, and Transparency*, ser. FAccT '22. ACM, Jun. 2022. [Online]. Available: http://dx.doi.org/10.1145/3531146.3533113

[19] M. S. A. Lee and J. Singh, "The landscape and gaps in open source fairness toolkits," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, ser. CHI '21. New York, NY, USA: Association for Computing Machinery, 2021. [Online]. Available: https://doi.org/10.1145/3411764.3445261

[20] S. Lambiase, G. Catolino, F. Palomba, F. Ferrucci, and D. Russo, "Investigating the role of cultural values in adopting large language models for software engineering," *arXiv preprint arXiv:2409.05055*, 2024.

[21] K. Tamilmani, N. Rana, S. Fosso Wamba, and R. Dwivedi, "The extended unified theory of acceptance and use of technology (utaut2): A systematic literature review and theory evaluation," *International Journal of Information Management*, vol. 57, p. 102269, 11 2020.

[22] B. Rakova, J. Yang, H. Cramer, and R. Chowdhury, "Where responsible ai meets reality: Practitioner perspectives on enablers for shifting organizational practices," *Proc. ACM Hum.-Comput. Interact.*, vol. 5, no. CSCW1, apr 2021. [Online]. Available: https://doi.org/10.1145/3449081

[23] V. Venkatesh, J. Y. L. Thong, and X. Xu, "Consumer acceptance and use of information technology: Extending the unified theory of acceptance and use of technology," *MIS Quarterly*, vol. 36, no. 1, pp. 157–178, 2012.

[24] J. F. Hair Junior, G. T. M. Hult, C. M. Ringle, and M. Sarstedt, "A primer on partial least squares structural equation modeling (pls-sem)," 2014.

[25] C. Starke, J. Baleis, B. Keller, and F. Marcinkowski, "Fairness perceptions of algorithmic decision-making: A systematic review of the empirical literature," 2021.

[26] S. Verma and J. Rubin, "Fairness definitions explained," in *2018 ieee/acm international workshop on software fairness (fairware)*. IEEE, 2018, pp. 1–7.

[27] S. Majumder, J. Chakraborty, G. R. Bai, K. T. Stolee, and T. Menzies, "Fair enough: Searching for sufficient measures of fairness," *ACM Trans. Softw. Eng. Methodol.*, vol. 32, no. 6, sep 2023. [Online]. Available: https://doi.org/10.1145/3585006

[28] S. Sharma, Y. Zhang, J. M. Ríos Aliaga, D. Bouneffouf, V. Muthusamy, and K. R. Varshney, "Data augmentation for discrimination prevention and bias disambiguation," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 2020, pp. 358–364.

[29] F. Calmon, D. Wei, B. Vinzamuri, K. Natesan Ramamurthy, and K. R. Varshney, "Optimized pre-processing for discrimination prevention," *Advances in neural information processing systems*, vol. 30, 2017.

[30] J. Chakraborty, S. Majumder, and T. 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.

[31] B. H. Zhang, B. Lemoine, and M. Mitchell, "Mitigating unwanted biases with adversarial learning," in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 2018, pp. 335–340.

[32] T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma, "Fairness-aware classifier with prejudice remover regularizer," in *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2012, Bristol, UK, September 24-28, 2012. Proceedings, Part II 23*. Springer, 2012, pp. 35–50.

[33] F. Kamiran and T. Calders, "Data preprocessing techniques for classification without discrimination," *Knowledge and information systems*, vol. 33, no. 1, pp. 1–33, 2012.

[34] J. Chakraborty, S. Majumder, Z. Yu, and T. Menzies, "Fairway: a way to build fair ml software," in *Proceedings of the 28th ACM joint meeting on European software engineering conference and symposium on the foundations of software engineering*, 2020, pp. 654–665.

[35] S. Galhotra, Y. Brun, and A. Meliou, "Fairness testing: testing software for discrimination," in *Proceedings of the 2017 11th Joint meeting on foundations of software engineering*, 2017, pp. 498–510.

[36] S. Udeshi, P. Arora, and S. Chattopadhyay, "Automated directed fairness testing," in *Proceedings of the 33rd ACM/IEEE international conference on automated software engineering*, 2018, pp. 98–108.

[37] Z. Chen, J. M. Zhang, F. Sarro, and M. Harman, "Fairness improvement with multiple protected attributes: How far are we?" in *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering*, 2024, pp. 1–13.

[38] C. Ferrara, F. Casillo, C. Gravino, A. De Lucia, and F. Palomba, "Refair: Toward a context-aware recommender for fairness requirements engineering," in *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering*, 2024, pp. 1–12.

[39] C. Ferrara, G. Sellitto, F. Ferrucci, F. Palomba, and A. De Lucia, "Fairness-aware machine learning engineering: how far are we?" *Empirical Software Engineering*, vol. 29, no. 1, p. 9, 2024.

[40] S. Vasudevan and K. 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] J. M. Zhang and M. Harman, ""ignorance and prejudice" in software fairness," in *2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE)*. IEEE, 2021, pp. 1436–1447.

[42] E. Sesari, F. Sarro, and A. Rastogi, "Understanding fairness in software engineering: Insights from stack exchange," *arXiv preprint arXiv:2402.19038*, 2024.

[43] G. Voria, G. Sellitto, C. Ferrara, F. Abate, A. De Lucia, F. Ferrucci, G. Catolino, and F. Palomba, "A catalog of fairness-aware practices in machine learning engineering," *arXiv preprint arXiv:2408.16683*, 2024.

[44] ——, "Fairness-aware practices from developers' perspective: A survey," *Available at SSRN 4949224*, 2024.

[45] A. Balayn, M. Yurrita, J. Yang, and U. Gadiraju, ""fairness toolkits, a checkbox culture?" on the factors that fragment developer practices in handling algorithmic harms," in *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, ser. AIES '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 482–495. [Online]. Available: https://doi.org/10.1145/3600211.3604674

[46] J. Wexler, M. Pushkarna, T. Bolukbasi, M. Wattenberg, F. Viégas, and J. Wilson, "The what-if tool: Interactive probing of machine learning models," *IEEE transactions on visualization and computer graphics*, vol. 26, no. 1, pp. 56–65, 2019.

[47] "scikit-fairness," 2019. [Online]. Available: https://github.com/koaning/scikit-fairness

[48] K. Holstein, J. Vaughan, I. Daumé, H., M. Dudík, and H. Wallach, "Improving fairness in machine learning systems: What do industry practitioners need?" Association for Computing Machinery, 2019, cited By 214.

[49] V. Venkatesh and M. Morris, "Why don't men ever stop to ask for directions? gender, social influence, and their role in technology acceptance and usage behavior," *MIS Quarterly*, vol. 24, pp. 115–139, 03 2000.

[50] V. Venkatesh, M. G. Morris, G. B. Davis, and F. D. Davis, "User acceptance of information technology: Toward a unified view," *MIS quarterly*, pp. 425–478, 2003.

[51] A. Momani, "The unified theory of acceptance and use of technology: A new approach in technology acceptance," *International Journal of Sociotechnology and Knowledge Development*, vol. 12, pp. 79–98, 07 2020.

[52] F. D. Davis, R. Bagozzi, and P. Warshaw, "Technology acceptance model," *J Manag Sci*, vol. 35, no. 8, pp. 982–1003, 1989.

[53] A. Authors, "Online appendix," 2024. [Online]. Available: https://figshare.com/s/583036589ad8b8abf666

[54] M. Figueroa-Armijos, B. Clark, and S. Veiga, "Ethical perceptions of ai in hiring and organizational trust: The role of performance expectancy and social influence," *Journal of Business Ethics*, vol. 186, 06 2022.

[55] L. Oneto and S. Chiappa, *Fairness in Machine Learning*. Cham: Springer International Publishing, 2020, pp. 155–196.

[56] F. D. Davis, "Perceived usefulness, perceived ease of use, and user acceptance of information technology," *MIS quarterly*, pp. 319–340, 1989.

[57] H. Van der Heijden, "User acceptance of hedonic information systems," *MIS quarterly*, pp. 695–704, 2004.

[58] F. Faul, E. Erdfelder, A. Buchner, and A.-G. Lang, "Statistical power analyses using g* power 3.1: Tests for correlation and regression analyses," *Behavior research methods*, vol. 41, no. 4, pp. 1149–1160, 2009.

[59] B. A. Kitchenham and S. L. Pfleeger, "Personal opinion surveys," in *Guide to advanced empirical software engineering*. Springer, 2008, pp. 63–92.

[60] D. Andrews, B. Nonnecke, and J. Preece, "Conducting research on the internet:: Online survey design, development and implementation guidelines," 2007.

[61] D. Russo and K.-J. Stol, "Pls-sem for software engineering research: An introduction and survey," *ACM Computing Surveys (CSUR)*, vol. 54, no. 4, pp. 1–38, 2021.

[62] A. Danilova, A. Naiakshina, S. Horstmann, and M. Smith, "Do you really code? designing and evaluating screening questions for online surveys with programmers," in *2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE)*. IEEE, 2021, pp. 537–548.

[63] P. Ralph, N. b. Ali, S. Baltes, D. Bianculli, J. Diaz, Y. Dittrich, N. Ernst, M. Felderer, R. Feldt, A. Filieri *et al.*, "Empirical standards for software engineering research," *arXiv preprint arXiv:2010.03525*, 2020.

[64] T. Hall and V. Flynn, "Ethical issues in software engineering research: a survey of current practice," *Empirical Software Engineering*, vol. 6, no. 4, pp. 305–317, 2001.

[65] C. M. Ringle, S. Wende, and J.-M. Becker, "Smartpls 4," Bönningstedt, 2024. [Online]. Available: {https://www.smartpls.com/}

[66] J. Henseler, C. M. Ringle, and M. Sarstedt, "A new criterion for assessing discriminant validity in variance-based structural equation modeling," *Journal of the academy of marketing science*, vol. 43, pp. 115–135, 2015.

[67] W. W. Chin *et al.*, "The partial least squares approach to structural equation modeling," *Modern methods for business research*, vol. 295, no. 2, pp. 295–336, 1998.

[68] S. Raithel, M. Sarstedt, S. Scharf, and M. Schwaiger, "On the value relevance of customer satisfaction. multiple drivers and multiple markets," *Journal of the academy of marketing science*, vol. 40, pp. 509–525, 2012.

[69] G. Shmueli, S. Ray, J. M. V. Estrada, and S. B. Chatla, "The elephant in the room: Predictive performance of pls models," *Journal of business Research*, vol. 69, no. 10, pp. 4552–4564, 2016.

[70] G. Shmueli, M. Sarstedt, J. F. Hair, J.-H. Cheah, H. Ting, S. Vaithilingam, and C. M. Ringle, "Predictive model assessment in pls-sem: guidelines for using plspredict," *European journal of marketing*, vol. 53, no. 11, pp. 2322–2347, 2019.

[71] R. L. Thompson, C. A. Higgins, and J. M. Howell, "Personal computing: Toward a conceptual model of utilization," *MIS quarterly*, pp. 125–143, 1991.

[72] C. Wohlin, P. Runeson, M. Höst, M. C. Ohlsson, B. Regnell, A. Wesslén *et al.*, *Experimentation in software engineering*. Springer, 2012, vol. 236.