Evaluating Behavior Change Interventions for Responsible Data Science

Ziwei Dong
Emory University
Atlanta, GA, USA
ziwei.dong@alumni.emory.edu

Keke Wu
Emory University
Atlanta, GA, USA
keke.wu@emory.edu

Leilani Battle
University of Washington
Seattle, WA, USA
leibatt@cs.washington.edu

Emily Wall
Emory University
Atlanta, GA, USA
emily.wall@emory.edu

Abstract

The adoption of responsible data science (RDS) practices in AI development remains inadequate despite growing awareness of algorithmic harms. One measure of success is by observing practitioners' *behaviors* – namely, their adoption of responsible sequences of behaviors in their model building practice. This paper evaluates two interventions for changing problematic *behaviors*: (i) a motivational priming intervention that introduces short, relevant stories, and (ii) a fairness toolkit (Aequitas)—to bridge the gap between ethical principles and practitioner behavior. Through a mixed-methods study with data scientists (N=12), we assess how these interventions influence fairness practices, model outcomes, and cognitive load across credit risk and income classification tasks. Results indicate that both interventions were efficient in promoting responsible data science behaviors and improving the delivered models' fairness, while maintaining baseline accuracy. We argue that effective behavior change interventions must balance technical tooling with motivational scaffolding to provide actionable insights for fostering sustainable RDS practices.

CCS Concepts

• **Human-centered computing** → **User studies**; **Empirical studies in HCI**.

Keywords

Behavior Change, Responsible Data Science, Intervention Evaluation, User Study

ACM Reference Format:
Ziwei Dong, Keke Wu, Leilani Battle, and Emily Wall. 2026. Evaluating Behavior Change Interventions for Responsible Data Science. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3772318.3791566

CHI '26, Barcelona, Spain

ACM ISBN 979-8-4007-2278-3/2026/04
https://doi.org/10.1145/3772318.3791566

1 Introduction

As artificial intelligence (AI) and machine learning (ML) systems permeate critical domains such as healthcare[8], finance[7], and criminal justice[32], their societal impact becomes increasingly significant. While these technologies promise to streamline decision-making and enhance efficiency, they also pose ethical challenges[5]. Algorithmic biases, data inequities, and unintended consequences have drawn attention to the need for responsible data science (RDS) practices. For instance, research has documented cases where predictive models systematically discriminate against marginalized communities, leading to unequal access to opportunities such as housing or loans[31]. Responsible data science encompasses practices aimed at mitigating these harms, fostering fairness, and ensuring accountability in data-driven systems[1, 33, 42].

Efforts to advance fairness and accountability in data science have predominantly centered on outcome-oriented metrics, such as fairness indices, accuracy measures, and other performance benchmarks, as tools for evaluating model outputs[27, 29, 38]. These metrics are undeniably valuable as they quantify disparities and highlight areas for improvement, providing a starting point for addressing systemic issues. However, they fall short of capturing the broader context of how these outcomes are achieved. For instance, while a fairness metric may confirm that a model meets certain criteria for equitable predictions, it offers no insight into whether responsible practices—such as comprehensive data exploration, bias-aware modeling, or transparent reporting—were followed during development[9, 19]. This exclusive focus on outcomes risks a superficial adherence to fairness goals, where technical metrics are optimized without addressing the behavioral and cognitive processes that underlie decision-making. Bridging this gap requires evaluation methods that not only measure end results but also assess the processes and practices that shape these results, ensuring that responsible behaviors are systematically embedded in data science workflows.

Recently, Dong et al.[15, 16] assert that responsible data science is inherently tied to the *behaviors* and practices of the data scientists themselves. They argue that while technical innovations—such as fairness auditing tools and bias mitigation algorithms—play a crucial role in advancing RDS, they cannot succeed in isolation. The ultimate impact of these tools depends on how data scientists incorporate them into their workflows and decision-making processes. This requires addressing human factors, such as awareness of ethical considerations, commitment to fairness, and motivation to adopt responsible practices, which collectively drive the success or failure of RDS initiatives. By framing RDS as a function of behavior, Dong et al. shift the focus from merely providing technical solutions to fostering a culture of responsibility among practitioners [16]. Such a behavioral lens recognizes that ethical data science is not achieved solely through tools and guidelines but through the deliberate actions and decisions of individuals. However, while Dong et al. lay a theoretical foundation [14, 16] of behavior change interventions, these theories have yet to be empirically evaluated in the context of RDS, a gap which we aim to address in this work.

Driving behavior change in responsible data science (RDS) necessitates a focus on the underlying cognitive and motivational factors that influence data scientists' decision-making processes. For instance, research in behavior change frameworks, such as the COM-B model [28], highlights that Capability, Opportunity, and Motivation are critical elements in fostering sustained behavioral shifts. For data scientists, these elements manifest as the ability to understand and apply fairness concepts (Capability), access to tools and resources that facilitate responsible practices (Opportunity), and the intrinsic or extrinsic drive to prioritize ethical considerations in their workflows (Motivation). Nevertheless, despite the existence of theoretical frameworks such as COM-B [28] for behavior change, their practical adoption in real-world workflows in the domain of data science has yet to be explored. Factors such as competing priorities, time constraints, and a lack of actionable guidance may hinder the integration of responsible practices into everyday tasks. Addressing this gap requires interventions that not only educate but also empower and motivate data scientists to incorporate responsibility as a core component of their workflows.

In this paper, we contribute a study with N=12 data scientists that aims to bridge the gap between theoretical aspirations and practical challenges in fostering responsible behavior in data science workflows. Our work complements existing descriptive research by providing, to our knowledge, the first controlled experimental evaluation of behavior change interventions in responsible data science, moving beyond characterizing existing practices to systematically measuring intervention effectiveness and underlying mechanisms. We evaluate the efficacy of behavior change interventions (BCIs) designed to encourage practices such as fairness-aware modeling, bias mitigation, and comprehensive data exploration. To achieve this, we compare three levels of intervention: (1) a control condition with no explicit guidance, (2) a fairness primer aimed at raising awareness of biases, and (3) the use of Aequitas, an open-source toolkit for auditing bias and assessing fairness. By assessing these interventions, we aim to uncover how different approaches influence both the behaviors of data scientists and the resulting technical outcomes. Additionally, we examine factors such as cognitive load and usability to ensure that the interventions are practical and sustainable for real-world adoption. This multifaceted evaluation provides actionable insights into how behavior change can be effectively fostered in data science practices.

Our study found that both interventions significantly increased responsible behaviors compared to the control condition, with Aequitas demonstrating superior effectiveness in improving fairness metrics ($p < 0.01$), while the motivational prime showed stronger motivational effects. Importantly, neither intervention compromised model accuracy. We also found that Aequitas imposed higher cognitive load, while personal connection to fairness scenarios enhanced intervention effectiveness, particularly among participants who could relate to disadvantaged groups.

In this work we make the following contributions:

1. **Empirical evaluation of behavior change interventions**: We provide the first systematic comparison of motivational priming versus technical tooling approaches for promoting responsible data science practices, demonstrating that both can effectively increase responsible behaviors.
2. **Mixed-methods framework for intervention evaluation**: We establish a comprehensive evaluation approach combining behavioral observation, fairness metrics, cognitive load assessment, and qualitative analysis that can guide future research in this domain.
3. **Cognitive load trade-offs in Behavior Change Intervention for RDS**: We quantify the cognitive burden associated with fairness tools, revealing that more effective interventions (Aequitas) may impose higher mental demands, informing future tool design.
4. **Role of personal connection in ethical motivation**: We demonstrate that practitioners' ability to relate to affected groups significantly enhances intervention effectiveness, suggesting design principles for more impactful behavior change strategies.

Our findings reveal actionable insights for designing effective behavior change interventions that balance technical capability, motivational support, and cognitive feasibility to foster sustainable responsible data science practices.

2 Related Work

2.1 Responsible Data Science

Responsible Data Science has emerged as a multidisciplinary effort to address ethical challenges arising from the deployment of AI and machine learning systems in high-stakes domains such as healthcare[8], finance[7], and criminal justice[32]. Early work in RDS focused on developing technical frameworks to quantify and mitigate biases in algorithmic decision-making. For example, foundational tools like AI Fairness 360[4] introduced fairness metrics (e.g., demographic parity, equalized odds) and bias mitigation algorithms, enabling practitioners to audit models for disparities. Similarly, Feldman et al.[17] proposed methods to certify and remove disparate impact in datasets through pre-processing techniques, emphasizing outcome-oriented fairness—ensuring models meet predefined fairness criteria in their predictions. These tools have become critical for diagnosing and addressing biases in data and models, particularly as research highlights systemic discrimination in domains like loan approvals[31] and healthcare diagnostics[8].

However, critiques of RDS frameworks argue that an overreliance on technical solutions risks overlooking the human and organizational contexts in which these tools are applied. Dong et al.[15, 16], for instance, assert that RDS demands rigor in both technical responsibility *and* behavioral responsibility. Similarly, Crisan et al.[10] observed that many RDS methodologies neglect the behavioral and cognitive processes of data scientists, such as how ethical considerations are weighed against competing priorities like accuracy or efficiency. For instance, practitioners may use fairness toolkits to audit models post hoc but fail to integrate responsible practices proactively into their workflows due to time constraints or lack of incentives[31]. This gap is further exacerbated by the disconnect between technical fairness (e.g., optimizing for statistical parity) and procedural fairness—the extent to which responsible practices (e.g., bias-aware data exploration, transparency in reporting) are systematically embedded into workflows[1].

Recent scholarship calls for a paradigm shift in RDS, moving beyond outcome-centric metrics to address the behavioral roots of ethical decision-making. Human-centered approaches, such as those outlined by Aragon et al.[1], emphasize that responsible outcomes depend not only on algorithmic corrections but also on fostering a culture of accountability among practitioners. This includes encouraging critical reflection on data collection practices, iterative bias mitigation across the model lifecycle (pre-processing, in-processing, post-processing)[4, 17], and transparent communication of limitations to stakeholders[42]. Yet, as Purificato et al.[31] demonstrated in their analysis of loan approval systems, even when fairness tools are available, practitioners often deprioritize ethical considerations unless motivated or incentivized to do so.

This evolving discourse underscores a key insight: **technical tools alone cannot guarantee responsible outcomes without complementary interventions that target the behaviors, norms, and incentives shaping data science practices**. By framing RDS as both a technical and behavioral challenge, this study builds on prior work to evaluate how interventions can bridge the gap between theoretical fairness frameworks and their practical adoption.

2.2 Behavior Change Interventions in Data Science

Behavior Change Interventions (BCIs) in data science aim to address the human factors that hinder the adoption of responsible practices, bridging the gap between theoretical frameworks and real-world workflows. Grounded in behavioral psychology, the COM-B model[28] provides a robust theoretical foundation for designing such interventions. It posits that sustained behavioral change requires three interrelated components: Capability (knowledge and skills), Opportunity (environmental enablers), and Motivation (intrinsic/extrinsic drivers). Recent studies have operationalized this framework in the data science contexts[16], focusing on tools and strategies that target these dimensions.

Capability - focused interventions emphasize equipping data scientists with the skills to identify and mitigate biases. For example, Cruz et al.[11] introduced FairGBM, a fairness-aware gradient-boosting framework that simplifies the integration of fairness constraints during model training. Similarly, educational primers and workshops[31] have been used to raise awareness of ethical pitfalls, such as conflating correlation with causation in biased datasets. These approaches align with findings by Purificato et al.[31], who demonstrated that practitioners often lack actionable guidance on how to operationalize fairness principles without sacrificing model performance.

Opportunity - oriented interventions focus on reducing barriers to adopting responsible practices by embedding tools directly into data scientists' workflows. The Aequitas toolkit[22], for instance, provides interactive bias audits within computational notebooks, enabling real-time fairness assessments. Building on this, recent work by Wang et al.[40] in SuperNOVA: Design Strategies and Opportunities for Interactive Visualization in Computational Notebooks highlights how embedded visualization tools can lower the cognitive cost of responsible practices. By integrating interactive visualizations for bias detection and fairness metrics, tools like SuperNOVA make complex ethical considerations more accessible, encouraging proactive engagement during model development.

Motivation - remains the most challenging dimension to address. Recent empirical studies highlight systemic barriers to sustaining motivation. For example, Holstein et al.[21] conducted interviews with 35 industry practitioners and identified that even when fairness tools are available, practitioners lack institutional incentives to adopt them, particularly when ethical outcomes are not tied to performance evaluations. To counteract ethical dissonance, BCIs must connect technical decisions to tangible societal outcomes. Shen et al.[34] developed Value Cards, an educational toolkit that uses scenario-based reflection to help practitioners weigh trade-offs between accuracy and fairness. Participants exposed to these exercises reported heightened motivation to adopt responsible practices, as they better understood the human consequences of biased models. Similarly, D'Ignazio et al.[13] advocate for participatory design in data science, where collaboration with marginalized communities fosters ethical accountability by grounding abstract fairness metrics in real-world harms.

Despite progress, gaps persist. Most BCIs focus on isolated dimensions of COM-B rather than holistically addressing capability, opportunity, and motivation. This study advances this discourse by evaluating how complementary approaches—motivational priming and tool-based support—interact to drive behavioral shifts.

Furthermore, recent empirical work has begun to characterize how practitioners engage with fairness toolkits in practice. Lee and Singh[24] conducted a comprehensive landscape analysis of open source fairness toolkits, providing a comparative assessment of existing tools and identifying key gaps in the fairness toolkit ecosystem. Their systematic evaluation highlighted challenges in toolkit usability and practitioner adoption, reinforcing the need for empirical studies that examine not just toolkit availability but their actual effectiveness in changing practitioner behaviors. While their work provides valuable insights into the current state of fairness tools, it focuses on descriptive analysis rather than experimental evaluation of behavior change mechanisms. Moreover, Deng et al.[12] conducted think-aloud studies with industry practitioners to understand how they learn about and attempt to use fairness toolkits (Fairlearn[6] and AIF360[4]), identifying challenges around contextualization, communication with non-technical stakeholders, and time-constrained decision-making. Similarly, Balayn et al.[2] conducted interviews with ML practitioners to examine factors that fragment practices when using fairness toolkits, finding that toolkits can act as "double-edged swords"—increasing awareness of algorithmic fairness while potentially promoting a checkbox culture and narrowing focus away from broader harms.

These descriptive studies collectively establish that practitioners face significant barriers to adopting responsible practices, even when tools are available, and that awareness alone may not translate to behavioral change. However, these works focus on characterizing existing practices and identifying barriers rather than experimentally evaluating whether specific interventions can effectively change behaviors and improve outcomes. This gap between descriptive understanding and experimental validation represents a critical limitation in the field: while we know that prompting practitioners to consider ethics might intuitively lead to more ethical behavior, the magnitude, mechanisms, and sustainability of such effects remain empirically uncharacterized. **Our work addresses this limitation by providing, to our knowledge, the first controlled experimental evaluation of behavior change interventions in responsible data science**, measuring not just whether practitioners use tools, but whether interventions systematically improve responsible behaviors and fairness outcomes, and the mechanisms through which this change occurs. By rigorously quantifying intervention effects and comparing different approaches, we move beyond intuitive expectations to provide evidence-based insights for designing effective behavior change strategies in RDS.

3 Methods

We conducted a user study to evaluate behavior change interventions for responsible data science. The method outlined here is designed to address several gaps in the literature. For instance, we build upon theoretical contributes in behavior change for RDS [15, 16] by providing empirical support for the efficacy of behavior change interventions. Likewise, our study directly addresses the limitations of outcome-centric approaches by systematically observing the processes and behaviors that lead to responsible outcomes. Rather than solely measuring fairness metrics post hoc, we evaluate the entire workflow—from data exploration through model development to bias auditing—to capture how interventions influence the behavioral and cognitive processes underlying ethical decision-making. Finally, our approach responds to critiques by Dong et al. [15, 16] and Crisan et al. [10] that highlight the neglect of human factors in RDS methodologies. Our experimental design operationalizes the COM-B framework [28] by testing interventions that target different behavior change mechanisms. By comparing these approaches against a control condition, we can isolate specific mechanisms through which behavior change occurs, moving beyond the technical tool provision that dominates current RDS research toward a more nuanced understanding of how to foster sustainable responsible practices in real-world data science workflows. This research study was reviewed and approved by a university institutional review board (IRB). This section presents an overview of the study objectives, the tasks that participants were asked to perform, and the chosen interventions.

3.1 Tasks & Interventions

The study design (depicted in Figure 1) is divided into two within-subjects tasks:

(1) **Credit**: In this task, participants were asked to predict the risk of bank loans using the German Credit Dataset[20]. The task required participants to evaluate how features such as income, credit history, and employment status might influence loan decisions.
(2) **Census**: This task focused on predicting income brackets using the Census (Adult Income) dataset[3] to advise the state government to allocate low-income housing benefit aid. Participants had to consider factors such as education, occupation, and marital status, analyzing their impact on income predictions.

The order of the two tasks is randomized. Additionally, there are three conditions corresponding to three levels of intervention:

(1) **Control** (no intervention): Participants complete the tasks in a basic Google Colab notebook.
(2) **Prime**: Task instructions are augmented with a short story to prime participants to be mindful of model fairness. For example, we added the following text in a cell at the beginning of the Google Colab notebook for the Credit Task: *"As banks increasingly deploy artificial intelligence tools to make credit decisions, they are having to revisit an unwelcome fact about the practice of lending: Historically, it has been riddled with biases against protected characteristics, such as race, gender, and sexual orientation. Such biases are evident in institutions' choices in terms of who gets credit and on what terms. Research shows that 56% of the female applicants evaluated would have received an unfair offer compared to their male peers with worse credit profiles and less profitable, but otherwise similar, businesses. If biases played no role in credit decision-making, that percentage would have been zero."*
(3) **Aequitas** [22]: Participants use the Aequitas notebook plugin to complete the task. Aequitas is an open-source tool developed to audit data science models for bias and fairness. It provides data scientists interventions to assess model fairness and mitigate biases in predictive models. Within the design space introduced by Dong et al. [15], Aequitas exemplifies a behavior change intervention that enhances both Capability and Motivation within the COM-B model [28] (What) for experienced data scientists (Who). It functions as an in-situ tool during modeling workflows (Where) that operates at the post-processing of data science pipeline (When), using automated fairness auditing with visual feedback (How) to identify and mitigate bias while balancing fairness and performance metrics (Why).

Each participant completed both the Credit and Census modeling tasks in a counterbalanced design. The first task was randomly assigned between Credit or Census, and the assignment of tasks to conditions was counterbalanced across participants. This ensured that task order (Credit-first vs. Census-first) and task-condition pairings were balanced between the Prime and Aequitas groups.

We selected Prime and Aequitas to represent two fundamental approaches to behavior change in RDS that map to distinct dimensions of the COM-B framework [28]. Prime exemplifies motivation-focused interventions that use narrative framing and awareness-raising to influence ethical decision-making—an approach grounded in social psychology research showing that personal relevance and emotional connection drive behavioral change [26]. This represents a broader class of "soft" interventions that aim to shift practitioners' mindsets without providing technical capabilities. This condition also enables us to check our fundamental assumption that analysts will behave more responsibly when prompted to do so. The interesting question for this condition is whether a soft behavioral nudge is sufficient to yield responsible analysis outcomes. In contrast, Aequitas represents capability and opportunity-focused interventions that provide concrete technical tools embedded directly into workflows. This approach assumes that practitioners are motivated but lack the means to operationalize fairness. Aequitas exemplifies a broader class of "hard" interventions including bias measurement and visualization that reduce technical barriers to responsible practices. We specifically chose Aequitas over alternatives (e.g., AI Fairness 360 [4]) because it integrates seamlessly into Jupyter notebooks, a common data science environment, and it focuses on post-processing auditing rather than pre-processing or in-processing constraints, allowing us to observe how practitioners respond to fairness feedback after initial model development. This design choice allows us to evaluate not just whether tools help, but how practitioners engage with fairness information when it's presented after they've already invested effort in model development.

3.2 Participants

We recruited 12 participants, each with at least 5 years of experience in data science. These participants were recruited through direct recruitment emails and message invitations on LinkedIn to data science practitioners in the industry. To avoid potential priming effects, the recruitment messages did not mention–explicitly or implicitly–that this study focuses on responsible data science or data science rigor. Similarly, our screening survey focused on demographic information and did not ask participants about their prior experience with responsible data science. Participants were incentivized with a digital gift card worth $25 per hour for their participation, which included their time spent in both phases of the study, as well as follow-up surveys and interviews. We introduce the participants' demographic information within the Table 1.

3.3 Procedure

The study was conducted via Zoom, with each of the two tasks taking approximately one hour and separated by a minimum of 24 hours. Each of the two tasks were conducted on separate days for two reasons: (1) to reduce fatigue and (2) to minimize the risk of participants relying on short-term memory or learning effects from earlier tasks that could influence their behavior in subsequent tasks. Participants were asked to turn on screen sharing during the session. The Zoom meeting was screen and audio recorded to capture participant behaviors during data science practices. Participants were provided with a Google Colab Notebook containing dataset characteristics and task instructions corresponding to their condition in the study. All AI features including Gemini were disabled. All participants provided verbal responses to our interview questions, with the exception of one participant (P7) who shared their responses via email due to recent health problems that affected their voice.

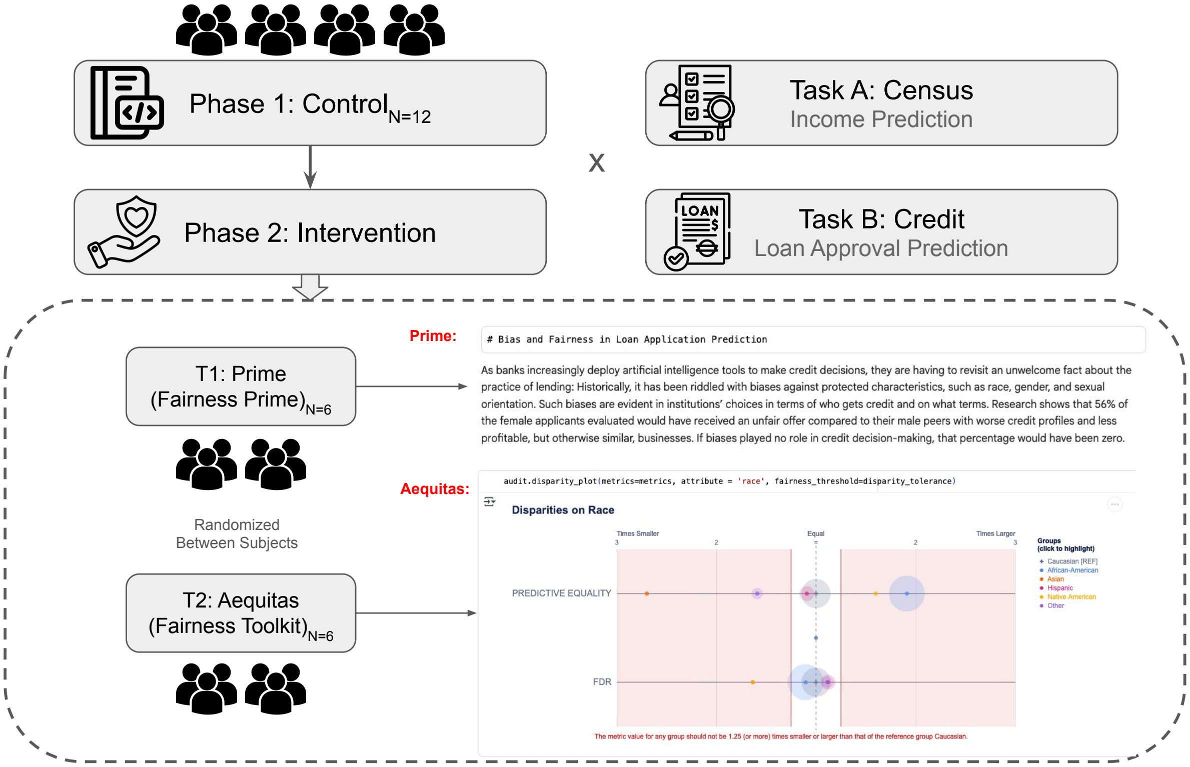
3.4 Responsible Data Science Practices and Data Collection

To evaluate the efficacy of behavior change interventions, we define an operational checklist of responsible data science practices. This checklist serves as a reference for observing participants' behaviors during their data science workflows. The practices identified are informed by Crisan et al.[10], which outlines the data science process, and Dong et al.[16], which provides concrete methods to ensure fairness and responsibility in AI-driven models. Below, we detail the specific responsible practices included in the checklist, categorized into *Pre-processing*, *In-processing*, and *Post-processing* stages.

(1) **Pre-processing -** These practices focus on preparing the data responsibly, ensuring quality and fairness before any modeling occurs.
   (a) Data Profiling: Examine data distributions and key attributes, such as column types and ranges, to identify anomalies or imbalances.
   (b) Data Engineering: Address missing values or labels in datasets by applying appropriate techniques such as imputation or removal.
   (c) Data Wrangling: Transform raw data into usable formats, such as converting numerical/textual data into categorical formats and creating meaningful features.
   (d) Sub-group Exploration: Analyze and visualize the distribution of sub-groups, particularly sensitive groups, to explore data distribution and potential data imbalances.
   (e) Bias Mitigation Pre-processing: Apply pre-processing bias mitigation methods, such as Disparity Impact Remover[17], to reduce bias in the data before modeling.

(2) **In-processing -** These practices guide responsible decision-making during model training.
   (a) Fairness-Aware Modeling: Utilize fairness-aware algorithms, such as FairGBM[11], to address biases during model development.
   (b) Compare Model Alternatives: Experiment with various model types to identify those that balance performance and fairness effectively.
   (c) Parameter Tuning: Perform hyperparameter comparison and optimization, such as grid or random search, to enhance model performance.

(3) **Post-processing -** These practices ensure the outputs of the model are interpreted and refined responsibly.
   (a) Outcome Inspection: Print or visualize confusion matrices or other evaluation metrics to understand model performance.
   (b) Bias Auditing: Audit the model's predictions to assess bias and fairness on the potentially disadvantaged group.
   (c) Configuration Iterations: Iterate on earlier stages (pre-processing or in-processing) based on the outcomes to refine the data or model configuration.
   (d) Fairness Correction: Modify model outputs to meet fairness criteria without retraining, such as equalizing outcomes to reduce disparities[30].



```
# Bias and Fairness in Loan Application Prediction
```

As banks increasingly deploy artificial intelligence tools to make credit decisions, they are having to revisit an unwelcome fact about the practice of lending: Historically, it has been riddled with biases against protected characteristics, such as race, gender, and sexual orientation. Such biases are evident in institutions' choices in terms of who gets credit and on what terms. Research shows that 56% of the female applicants evaluated would have received an unfair offer compared to their male peers with worse credit profiles and less profitable, but otherwise similar, businesses. If biases played no role in credit decision-making, that percentage would have been zero.

```
audit.disparity_plot(metrics=metrics, attribute = 'race', fairness_threshold=disparity_tolerance)
```

The metric value for any group should not be 1.25 (or more) times smaller or larger than that of the reference group Caucasian.

Figure 1: Participants all completed either Task A or B (order randomized) first with the Control condition. Then, participants were randomly assigned to either the Fairness Prime condition or the Aequitas condition to complete the other task.

Table 1: The participants' demographic information, including their gender, years of experience (YOE), age, and occupation

Session	Participant	Gender	YOE	Age	Occupation
Prime	P1	Male	5	26	Data Scientist
Prime	P2	Female	6	38	Data Scientist
Prime	P3	Female	11	37	Applied Scientist
Prime	P4	Male	7	38	Research Scientist
Prime	P5	Male	8	32	Data Scientist
Prime	P6	Female	9	31	Applied Scientist
Aequitas	P7	Male	8	31	Data Scientist
Aequitas	P8	Male	7	32	Bio Info Data Scientist
Aequitas	P9	Male	5	27	Data Analytist
Aequitas	P10	Male	10	36	Machine Learning Scientist
Aequitas	P11	Male	10	33	Data Scientist
Aequitas	P12	Male	8	36	Applied Scientist

Using this checklist, we can then derive a score for a given data science workflow according to how many of these responsible practices are observed.

3.5 Hypotheses & Measures

In this user study, we formulated 5 hypotheses regarding the impact of behavior change interventions on responsible data science practices. These hypotheses align with our paper's central goal of evaluating how behavior change interventions can bridge the gap between ethical principles and practitioner behavior in responsible data science. Through hypothesis testing, we aim to provide empirical evidence for designing effective interventions that promote sustainable responsible practices without compromising technical outcomes.

For each hypothesis (H1–H5), statistical tests (e.g., repeated-measures t-tests, Wilcoxon test) are applied to assess differences across intervention conditions (Control, Prime, Aequitas). We detail the hypotheses and respective statistical measurement for each hypothesis below:

H1: Responsible Behaviors *- Both Prime and Aequitas interventions will promote responsible behaviors compared to the Control, and Aequitas will outperform Prime.* We conduct a Repeated-Measures t-test to compare the mean coverage and frequency of responsible practices observed in participants' behaviors across the three intervention levels: Control, Prime, and Aequitas.

H2: COM-B Factors*[28] - Both Prime and Aequitas interventions will primarily influence the 'Motivation' factor within the COM-B framework, compared to 'Capability' and 'Opportunity'.* We utilize a Wilcoxon test to examine differences in participants' Likert-scale ratings for "Capability," "Opportunity," and "Motivation" across the three conditions, while accounting for within-subject dependencies.

H3: Model Fairness *- Both Prime and Aequitas interventions will improve fairness metrics of the resulting models, and Aequitas will outperform Prime.* We perform a repeated-measures t-test to compare fairness metrics (false discovery rate ratio of disadvantaged and non-disadvantaged groups) across conditions (Control, Prime, and Aequitas).

H4: Model Performance *- Both Prime and Aequitas interventions will not significantly affect the accuracy of the resulting models.* We perform a repeated-measures t-test to compare model's accuracy across conditions (Control, Prime, and Aequitas).

H5: Cognitive Load *- Both Prime and Aequitas interventions will not significantly increase the cognitive load compared to the Control.* We apply a Wilcoxon test to compare the median cognitive load scores between conditions (e.g., Control vs. Prime, Control vs. Aequitas) by analyzing the direction and magnitude of paired differences.

Rationale: We predict Aequitas will outperform Prime (H1, H3) because interventions addressing multiple COM-B dimensions (Capability + Motivation + Opportunity) produce stronger effects than single-dimension approaches (Motivation alone) [37], and actionable tools enable practitioners to act on fairness concerns rather than merely being aware of them [39]. We hypothesize Motivation as the primary driver for both interventions (H2) because our experienced participants (5+ years) already possess baseline technical skills, and the one-hour timeframe limits deep capability development. Instead, interventions are likely to shift how practitioners prioritize fairness relative to other goals. We predict no significant impacts to model accuracy (H4) because fairness and accuracy are not inherently opposed in this experiment [25], challenging the common "fairness-accuracy tradeoff" belief that hinders RDS adoption. Finally, we hypothesize minimal cognitive load increases (H5) as Prime requires only reading a brief narrative and Aequitas automates manual fairness auditing, though understanding cognitive costs is critical for real-world adoption [41].

4 Quantitative Findings

4.1 H1: Responsible Behaviors

To assess the impact of our interventions on responsible data science practices (**H1**), we counted the number of responsible behaviors exhibited by participants across all three conditions, using the framework outlined in subsection 3.4 (maximum of 12 responsible behaviors).

The data reveal a clear pattern of a larger number of responsible behaviors observed while using the interventions compared to the control. When exposed to the Prime intervention (Figure 2), participants exhibited an average of $\mu = 5.7$ ($\sigma = 0.75$) responsible behaviors compared to their Control behaviors of $\mu = 2.8$ ($\sigma = 0.69$). A repeated measures t-test revealed a large positive effect ($t = -9.21$, $p < 0.01$, $Cohen'sd$ [23] = 2.56, 0.95 CI (confidence interval [18]) [0.39, 4.72]). With this large effect size, we had >0.99 statistical power to detect the intervention's impact, and the confidence interval suggests the true effect ranges from medium to very large positive effects on responsible behaviors.

Those using the Aequitas toolkit (Figure 3) demonstrated the highest average at $\mu = 8.1$ ($\sigma = 1.34$) responsible behaviors compared to their Control behaviors of $\mu = 3.67$ ($\sigma = 1.25$), showing a large positive effect ($t = -6.26$, $p < 0.01$, $Cohen'sd = 3.76$, 0.95 CI [0.78, 6.75]). The confidence interval indicates the true effect ranges from small to very large, with >0.99 statistical power to detect this substantial intervention impact.

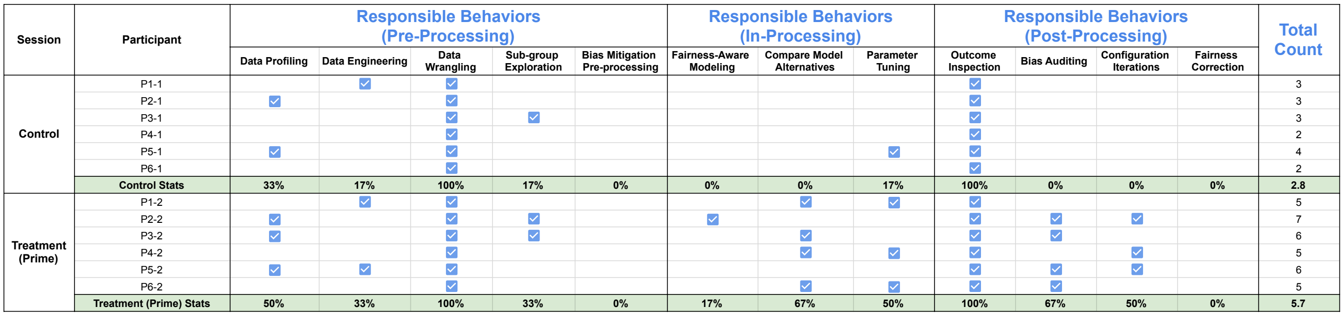
This observation suggested a marginally significant difference in baseline behaviors between the two groups. While this difference did not reach the conventional threshold for statistical significance ($p < 0.05$), it indicates some initial variability between groups that should be considered when interpreting the intervention effects.

These results **support H1**: that both interventions promote responsible behaviors compared to the Control condition, with Aequitas demonstrating a more pronounced effect than Prime. The significant increase in observed responsible behaviors suggests that both motivational priming and technical tooling such as Aequitas can effectively influence data scientists' behaviors towards more responsible practices.

These quantitative results are further supported by participants' reflections on their behavior during the study. For example, P6 noted, *"I definitely put more effort within this study to ensure fairness [after reading the Prime]."* Likewise, P3 noted, *"This tool made me aware how unfair some groups can be treated in ML practices… it gives me a chance to revisit my model configuration from time to time."* Similarly, participants using Aequitas described shifts in their practices. P1 reflected, *"Using Aequitas made me realize it [fairness] is also a very important component in the evaluation process… Without Aequitas, I would feel reluctant to put so much effort to manually implement and evaluate fairness on my own."*

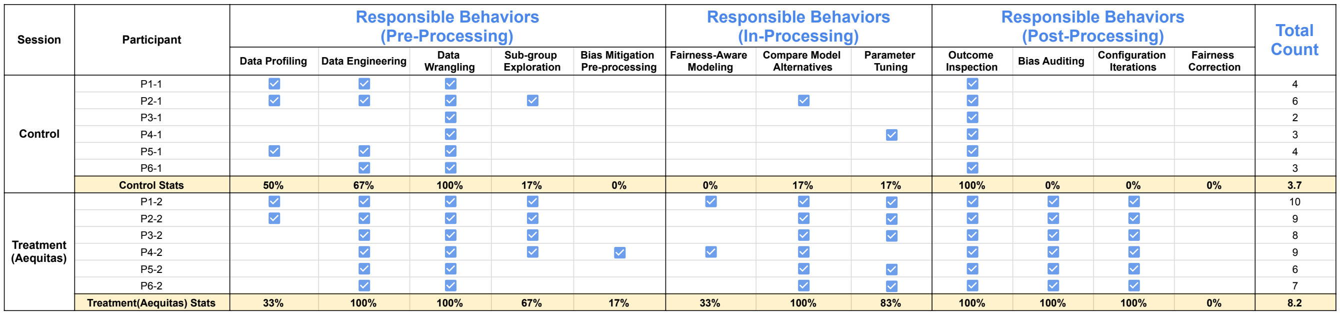
4.2 H2: COM-B Factors

For **H2**, we were interested in whether the interventions would primarily influence an individual's Capability (C), Opportunity (O), or Motivation (M) to perform responsible data science behaviors. We asked participants to rate the extent to which the treatment altered their Capability, Opportunity, and Motivation to complete the task responsibly, and asked them follow up questions to further elaborate on their rating (Interview questions are shown in Table 2). We used the term "engagement" to replace the term "opportunity" as pilot studies revealed that the terminology of "opportunity" was not clear to participants. We collected Likert-style ratings (-2 = Negative Influence, -1 = Slightly Negative Influence, 1 = Slightly



Session	Participant	Responsible Behaviors (Pre-Processing)					Responsible Behaviors (In-Processing)			Responsible Behaviors (Post-Processing)				Total Count
		Data Profiling	Data Engineering	Data Wrangling	Sub-group Exploration	Bias Mitigation Pre-processing	Fairness-Aware Modeling	Compare Model Alternatives	Parameter Tuning	Outcome Inspection	Bias Auditing	Configuration Iterations	Fairness Correction	
Control	P1-1		☑	☑						☑				3
	P2-1	☑		☑						☑				3
	P3-1			☑	☑					☑				3
	P4-1			☑						☑				2
	P5-1	☑		☑					☑	☑				4
	P6-1			☑						☑				2
	Control Stats	33%	17%	100%	17%	0%	0%	0%	17%	100%	0%	0%	0%	2.8
Treatment (Prime)	P1-2		☑	☑				☑	☑	☑				5
	P2-2	☑		☑	☑		☑			☑	☑	☑		7
	P3-2	☑		☑	☑			☑		☑	☑			6
	P4-2			☑				☑	☑	☑		☑		5
	P5-2	☑	☑	☑						☑	☑	☑		6
	P6-2			☑				☑	☑	☑	☑			5
	Treatment (Prime) Stats	50%	33%	100%	33%	0%	17%	67%	50%	100%	67%	50%	0%	5.7

Figure 2: An overview of the participants' responsible behaviors within the Prime group's control and treatment sessions.



Session	Participant	Responsible Behaviors (Pre-Processing)					Responsible Behaviors (In-Processing)			Responsible Behaviors (Post-Processing)				Total Count
		Data Profiling	Data Engineering	Data Wrangling	Sub-group Exploration	Bias Mitigation Pre-processing	Fairness-Aware Modeling	Compare Model Alternatives	Parameter Tuning	Outcome Inspection	Bias Auditing	Configuration Iterations	Fairness Correction	
Control	P1-1	☑	☑	☑						☑				4
	P2-1	☑	☑	☑	☑			☑		☑				6
	P3-1			☑						☑				2
	P4-1			☑					☑	☑				3
	P5-1	☑	☑	☑						☑				4
	P6-1		☑	☑						☑				3
	Control Stats	50%	67%	100%	17%	0%	0%	17%	17%	100%	0%	0%	0%	3.7
Treatment (Aequitas)	P1-2	☑	☑	☑	☑		☑	☑	☑	☑	☑	☑		10
	P2-2	☑	☑	☑	☑			☑	☑	☑	☑	☑		9
	P3-2		☑	☑	☑			☑	☑	☑	☑	☑		8
	P4-2		☑	☑	☑	☑	☑	☑		☑	☑	☑		9
	P5-2		☑	☑				☑	☑	☑	☑	☑		6
	P6-2		☑	☑				☑	☑	☑	☑	☑		7
	Treatment(Aequitas) Stats	33%	100%	100%	67%	17%	33%	100%	83%	100%	100%	100%	0%	8.2

Figure 3: An overview of the participants' responsible behaviors within Aequitas group's control and treatment sessions.

Positive Influence, 2 = Positive Influence) from participants on their perceived Capability, Opportunity, and Motivation across the three conditions (Table 2). We employed Wilcoxon signed rank tests to analyze differences between COM-B factors within each intervention group.

For the Prime intervention (n=6), participants rated a marginally positive influence on Capability ($\mu = 1$), Opportunity ($\mu = 1.67$), and Motivation ($\mu = 2$). Consistent with **H2**, Motivation was rated as having the largest influence. A Wilcoxon signed rank test comparing Motivation and Capability yielded a significant effect ($W = 21$, $p = 0.0715$, rank-biserial $r = 1.00$, indicating a large effect size), suggesting Motivation ratings were higher. However, the comparison of Motivation to Opportunity showed a large effect size but was not statistically significant ($W = 3$, $p = 0.079$, $r = 0.86$), likely due to limited statistical power with our small sample. For the Aequitas intervention (n=6), participants rated a marginally positive influence on Capability ($\mu = 1.67$), Opportunity ($\mu = 1.67$), and Motivation ($\mu = 2$). While Motivation was rated higher than both Capability ($W = 3$, $p = 0.079$, $r = 0.86$) and Opportunity ($W = 3$, $p = 0.079$, $r = 0.86$), the results were not statistically significant despite large effect sizes. The combination of large effect sizes ($r = 0.86$) with non-significant p-values reflects our limited statistical power with n=6, suggesting the data is consistent with meaningful differences in COM-B factors that our study was underpowered to detect definitively. Thus these findings **partially support H2**.

While the Prime intervention significantly enhanced participants' Motivation compared to their perceived Capability, it was not significant when comparing Motivation to Opportunity, or for any of the factors for the Aequitas intervention. While we do not find conclusive statistical support for the motivational impact of the interventions, we do find qualitative support for some individuals. Several participants who received the Prime intervention explicitly mentioned how the fairness framing motivated their behaviors. For instance, P5 shared, *"I wouldn't necessarily say my model is better, but I would say it's more responsible given my motivation... It [Prime] definitely motivates me to explore more about what's going on with this bias."* Similarly, P10 noted, *"It made the problem more tangible, more personal, and more interesting to solve, and made me look forward more to the outcome of this data science model."* Participants exposed to Aequitas also reported motivational effects, though slightly more nuanced. P8 observed, *"It improved my motivation by making it visually prominent... bias is something small and implicit, not explicitly captured by any commonly used metrics in my usual workflow."* Thus while Capability and Opportunity may still played a role, participants largely perceived Motivation as the driver of behavior change.

The pattern of large effect sizes (r>0.8) coupled with non-significant p-values for several comparisons illustrates the importance of reporting effect sizes alongside significance tests. With our small sample size, we had limited power to detect differences in ordinal ratings, yet the effect size estimates suggest that Motivation may indeed be more strongly influenced than other COM-B factors, particularly for the Prime intervention where the Motivation vs Capability comparison reached statistical significance.

Table 2: Interview questions we asked participants after they finished the treatment study session.

Factor	Question Phrasing
C	How did Prime/Aequitas influence your *capability* to complete the task?
O	How did Prime/Aequitas influence your *engagement* to complete the task?
M	How did Prime/Aequitas influence your *motivation* to complete the task?

4.3 H3: Model Fairness

To evaluate the impact of the interventions on model fairness (**H3**), we performed repeated-measures t-tests on fairness metrics across the Control, Prime, and Aequitas conditions. We operationalized fairness as the false discovery rate ratio between disadvantaged and non-disadvantaged groups. In this case, values closer to 1 indicate more fair models. We hypothesized that both interventions would lead to fairness improvements, with Aequitas yielding superior results compared to Prime.

Participants in the Prime condition exhibited a slight improvement in fairness metrics ($\mu = 1.65$, $\sigma = 0.35$) over their Control fairness ($\mu = 2.05$, $\sigma = 0.39$). A paired t-test revealed a medium effect size ($t = 1.42$, $p = 0.1075$, $Cohen'sd = 0.58$, 0.95 CI [-0.55, 1.71]). The wide confidence interval reflects substantial uncertainty about the impact on model fairness. While not reaching conventional significance thresholds and with only 0.21 statistical power to detect medium effects given our sample size (n=6), the non-significant p-value suggests no strong evidence of fairness changes, though we cannot definitively rule out meaningful impacts in either direction. In contrast, participants in the Aequitas condition demonstrated a substantial improvement in fairness ($\mu = 1.18$, $\sigma = 0.16$) over their Control fairness ($\mu = 2.45$, $\sigma = 0.31$) , showing a very large effect ($t = 9.2$, $p < 0.01$, $Cohen'sd = 3.76$, 0.95 CI [0.78, 6.74]). With >0.99 statistical power to detect this large effect, the confidence interval strongly suggests substantial positive impacts on model fairness. This effect suggests that providing direct fairness auditing and mitigation tools can significantly impact fairness outcomes, leading to more equitable predictions across demographic groups. Given these findings, we find **partial support for H3**: Aequitas significantly improves fairness metrics, while Prime shows a trend toward improvement that did not reach statistical significance. Aequitas outperforms Prime in terms of fairness improvement. The quantitative improvements in model fairness were echoed by participants' reflections on how the interventions shaped their fairness-oriented decision-making. Participants who used Aequitas, for instance, highlighted the tool's role in identifying and mitigating bias. P1 shared, *"Using Aequitas made me realize fairness is also a very important component in the evaluation process... It can inform my decision on what kind of model to choose."*

4.4 H4: Model Performance

To evaluate whether the interventions affected model performance (**H4**), we conducted repeated-measures t-tests comparing the accuracy of models produced in the Control condition versus those developed with the Prime and Aequitas interventions. For the Prime group (n=6), models showed a mean accuracy of $\mu = 0.60$ ($\sigma = 0.03$) compared to Control accuracy of μ=0.59 (σ=0.05), representing a small positive effect ($t = -0.59$, $p = 0.58$, $Cohen'sd = 0.24$, 0.95 CI [-0.83, 1.30]). The Aequitas group (n=6) showed mean accuracy of μ=0.58 (σ=0.04) compared to Control performance of μ=0.61 (σ=0.02), representing a medium negative effect ($t = 1.59$, $p = 0.17$, $Cohen'sd$=-0.65, 0.95 CI [-1.80, 0.51]). Critically, given our sample size (n=6 per group), we had only 0.17 statistical power to detect medium-sized effects on accuracy. The wide confidence intervals reflect substantial uncertainty about the impact on model accuracy. While the non-significant p-values suggest no strong evidence of accuracy changes, we cannot rule out meaningful impacts in either direction. Thus, we interpret these results as indicating that our study was underpowered to detect such effects definitively.Future studies with larger sample sizes would be needed to narrow these confidence intervals and make more definitive claims about accuracy impacts.

These findings have important implications for the responsible data science field. We found that **H4 is non-significant (underpowered)**: The combination of small effect sizes and wide confidence intervals suggests that fairness improvements (as demonstrated in H3) need not come at a substantial cost to model performance. While we cannot definitively rule out small accuracy trade-offs, the magnitude of any such trade-offs appears modest relative to the substantial fairness gains observed, particularly for Aequitas. This challenges the common perception that fairness and accuracy exist in strict opposition, suggesting instead that with appropriate interventions, practitioners can improve fairness outcomes while maintaining acceptable model quality. We acknowledge that there exist cases where fairness and accuracy may be in tension—for instance, when targeted discrimination would improve overall model performance—but we believe these findings contribute to understanding contexts where responsible practices need not compromise technical outcomes.

4.5 H5: Cognitive Load

To assess **H5**, cognitive load was measured using the NASA-TLX on a 7-point scale (0 = low demand to 7 = high demand) for six dimensions: mental demand, physical demand, temporal demand, performance, effort, and frustration. Note that because this was a non-standard scale for the NASA-TLX, we rely on within-subjects relative comparison of ratings, with each response first standardized before conducting statistical tests. We employed the Wilcoxon signed-rank test to determine whether the interventions significantly influenced cognitive load across the six dimensions of the NASA-TLX. Table 3 summarizes the standardized ratings for both groups and the corresponding p-values from the Wilcoxon test.

For the Prime intervention, we compared standardized NASA-TLX scores between the control and treatment groups. A Wilcoxon signed rank test revealed no statistically significant differences in mental demand ($\mu_C = 2.5$, $\mu_P = 2.67$, $W = 3.5$, $p = 0.39$), physical demand ($\mu_C = 0.67$, $\mu_P = 1.5$, $W = 3$, $p = 0.08$), temporal demand ($\mu_C = 0.5$, $\mu_P = 1.33$, $W = 3$, $p = 0.09$), performance ratings ($\mu_C = 4.17$, $\mu_P = 4$, $W = 2$, $p = 0.72$), effort ($\mu_C = 2$, $\mu_P = 2.3$, $W = 10$, $p = 0.24$), or frustration ($\mu_C = 0.83$, $\mu_P = 1.17$, $W = 1$, $p = 0.16$). Effect sizes (rank-biserial correlation) for the Prime group were generally large across dimensions (r ranging from 0.52 to 0.95), though the absolute magnitude of changes remained small (typically <1 point on the 7-point scale). These findings suggest that the Prime intervention did not introduce additional cognitive burden on participants. This aligns with the intervention's design as a lightweight, awareness-raising tool that required no procedural changes to participants' workflows.

In contrast, for the Aequitas intervention, the Wilcoxon signed-rank test indicated a statistically significant increase in mental demand ($\mu_C = 1.67$, $\mu_A = 3$, $W = 21$, $p = 0.02$), physical demand ($\mu_C = 0.67$, $\mu_A = 1.33$, $W = 10$, $p = 0.02$), performance ($\mu_C = 2.33$, $\mu_A = 3.33$, $W = 10$, $p = 0.03$), and effort ($\mu_C = 1$, $\mu_A = 2.83$, $W = 15$, $p = 0.02$) in the treatment group compared to the control. However, no significant differences were observed in temporal demand ($\mu_C = 1.17$, $\mu_A = 1.67$, $W = 4.5$, $p = 0.21$) or frustration ($\mu_C = 0.67$, $\mu_A = 0.67$, $W = 3$, $p = 0.5$). Effect sizes for Aequitas varied more substantially, with small effects for mental demand (r=0.00) and effort (r=0.29), and large effects for other dimensions (r ranging from 0.52 to 0.86). The significant increases in mental demand (+1.33 points) and effort (+1.83 points) suggest that while Aequitas effectively promotes fairness, it does so by requiring additional cognitive engagement from practitioners. These results suggest that while Aequitas may enhance fairness-oriented decision-making, it also imposes a higher cognitive load in specific areas. These results indicate that the toolkit's additional steps—such as configuring bias audits and interpreting fairness metrics—required more cognitive engagement than the Control or Prime conditions. We these findings **partially support H5**: Prime does not increase cognitive load, while Aequitas *does* increase cognitive load.

Three participants' perceived experiences align with our findings on cognitive load. Participants using Aequitas described a more involved and demanding experience. P7 noted, *"This tool increases my ability of doing data science work but not a lot, but I do think it almost mandatorily engaged myself on checking the fairness status,"* indicating increased effort and mental engagement. P1 echoed this, saying, *"Without Aequitas, I would feel reluctant to put so much effort to manually implement and evaluate the fairness by my own."* These comments suggest that while Aequitas was effective in promoting fairness, it did so by increasing the cognitive demands of the task.

4.6 Summary of Results

We summarize the outcomes for the hypotheses in Figure 4. Our results demonstrate that behavior change interventions can effectively promote responsible data science practices, with technical tools like Aequitas showing stronger impacts on fairness outcomes than motivational priming alone. However, this effectiveness comes with a cognitive cost, as the more effective intervention (Aequitas) also imposed higher cognitive demands. Importantly, we found that improving fairness did not necessarily require sacrificing model performance. These findings suggest that intervention designers should consider the balance between effectiveness and cognitive burden, potentially exploring hybrid approaches that combine motivational elements with streamlined technical capabilities.

5 Qualitative Findings

5.1 Methodology

To complement our quantitative findings, we conducted a qualitative analysis of participant interviews to understand the mechanisms through which behavior change interventions influence data scientists' practices and decision-making processes. This subsection describes our approach and method to analyzing the qualitative data collected during the study.

5.1.1 Quote Selection and Initial Coding. Our qualitative analysis began with transcription of all interviews. We aimed to identify valuable quotes that provided insights into participants' experiences, motivations, and behavioral changes. To develop our quote selection criteria, we conducted an initial review of all interview transcripts to understand the range and nature of participant responses. This preliminary analysis revealed that participants expressed diverse reactions to the interventions, including positive impacts, challenges encountered, neutral observations, and critical reflections. Based on this review, we established inclusion and exclusion criteria that would capture the breadth of participant experiences while focusing on quotes that provided meaningful insights into the mechanisms of behavior change.

Quotes were retained if they met one or more of the following inclusion criteria:

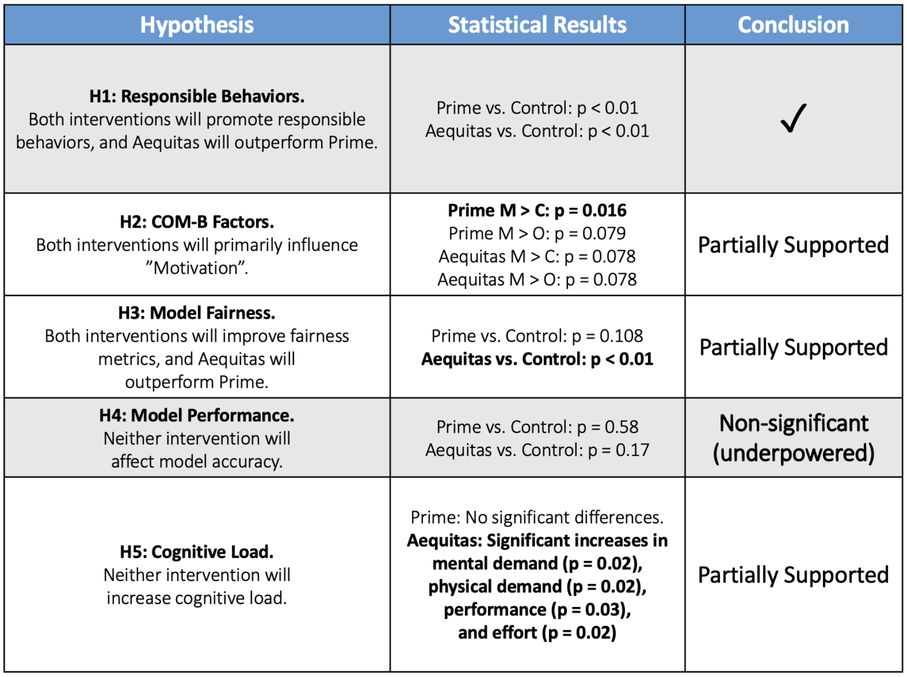
(1) explicit reflection on how the intervention influenced their approach, decision-making processes, or workflow compared to their usual practices, regardless of whether the influence was perceived as positive, negative, or mixed;
(2) concrete descriptions of specific behavioral changes or new practices adopted during the study;
(3) insights into underlying motivational factors, cognitive processes, or emotional responses triggered by the interventions;
(4) direct comparisons between their control session experience and treatment session experience, covering both improvements and difficulties;
(5) expressions of future intentions to adopt responsible data science practices or use fairness tools in their work, including both enthusiasm or concerns for adoptions;
(6) articulation of new understanding or awareness about fairness, bias, or responsible data science concepts.

This process resulted in 47 retained quotes from across the 12 participants, which were compiled in a structured spreadsheet for further analysis. Each quote was attributed to its participant and marked for use in the final analysis. This process ensured that we captured the most meaningful and representative insights from participants' experiences while maintaining the richness and nuance of their reflections.

5.1.2 Thematic Analysis and Clustering. Following the initial quote selection, we conducted a thematic analysis[36] to identify recurring patterns and group related insights into coherent themes. Using an iterative process, we conducted open coding on the selected

Table 3: Comparison of NASA-TLX cognitive load dimensions across intervention conditions (Control vs. Prime vs. Aequitas) using Wilcoxon signed-rank tests. We report the within-group average difference between control and treatment and indicate significance as p-values < 0.05 with an asterisk*.

Session	Mental	Physical	Temporal	Performance	Effort	Frustration
Prime	0.17	0.83	0.83	0	0.33	0.33
Aequitas	**1.33***	**0.33***	0.5	**1***	**1.83***	0



Hypothesis	Statistical Results	Conclusion
H1: Responsible Behaviors. Both interventions will promote responsible behaviors, and Aequitas will outperform Prime.	Prime vs. Control: p < 0.01 Aequitas vs. Control: p < 0.01	✓
H2: COM-B Factors. Both interventions will primarily influence "Motivation".	**Prime M > C: p = 0.016** Prime M > O: p = 0.079 Aequitas M > C: p = 0.078 Aequitas M > O: p = 0.078	Partially Supported
H3: Model Fairness. Both interventions will improve fairness metrics, and Aequitas will outperform Prime.	Prime vs. Control: p = 0.108 **Aequitas vs. Control: p < 0.01**	Partially Supported
H4: Model Performance. Neither intervention will affect model accuracy.	Prime vs. Control: p = 0.58 Aequitas vs. Control: p = 0.17	Non-significant (underpowered)
H5: Cognitive Load. Neither intervention will increase cognitive load.	Prime: No significant differences. **Aequitas: Significant increases in mental demand (p = 0.02), physical demand (p = 0.02), performance (p = 0.03), and effort (p = 0.02)**	Partially Supported

Figure 4: An overview of the outcomes for the 5 hypotheses we proposed, detailed statistical results and interpretations can be found in section 4.

quotes to identify preliminary concepts and patterns, then grouped similar quotes and concepts into clusters based on shared characteristics. We developed higher-level themes that captured the essence of each cluster and refined theme definitions through discussion and consensus between researchers. Finally, we validated themes by ensuring each theme is supported by multiple participant quotes and represents insights that emerged consistently across different participants and intervention conditions.

The final themes were selected based on their prevalence across participants, their relevance to understanding behavior change mechanisms, and their potential to inform future intervention design. This qualitative analysis approach allowed us to uncover nuanced insights into how behavior change interventions operate in practice, providing depth and context to complement our quantitative findings about intervention effectiveness. We ended up having five primary themes that extended beyond our hypothesis-driven quantitative analysis: (I) Foster Empathy, (II) Visualize Fairness for Understanding and Action, (III) Expand From the Focus of Traditional Metrics, (IV) Integrate into Existing DS Workflows, and (V) Boost Intrinsic Motivation. Each theme emerged organically from the data and represented consistent patterns observed across multiple participants and intervention conditions.

5.2 Theme I: Foster Empathy

A striking finding that emerged from our interviews was how personal connection to fairness issues significantly influenced participants' engagement with responsible practices. This was particularly evident among female participants who could directly relate to the scenarios presented in the interventions. As P3 reflected, "I can totally relate to the situation (female applicants being unfairly treated by DS models) as a female," while P6 similarly noted, "As a female, I can relate to that"..."It made me be aware of the disadvantage my gender is suffering (from), so I paid more attention to the fairness here." This personal connection appeared to motivate more thorough engagement with fairness considerations, as evidenced by P6's follow-up action: "Due to the Prime, I would spend more effort to try to learn a more unbiased model." P3 elaborated on this motivation, stating "After reading the Prime, I realized the model is actually made to (...) change people's life, so I would pay extra attention [to] making any data cleaning or model development choices." This finding suggests that future behavior change interventions could be more effective by incorporating personalized scenarios and real-world examples that help practitioners connect emotionally with the potential impacts of their work, particularly by highlighting experiences of affected groups that resonate with practitioners' own identities or lived experiences.

5.3 Theme II: Visualize Fairness for Understanding and Action

The visual representation of fairness metrics emerged as a crucial factor in helping participants understand and address bias. This theme was naturally prominent among Aequitas users, who frequently mentioned how visualization transformed their understanding of model bias. P8 noted, "It gives the awareness of model bias for the first time, and I can visually see how biased my model outcome is using this tool," while P12 similarly reflected, "After I can visually see bias of my model towards a feature, it made me more concerned and careful about bias that can be introduced in my model." The visual feedback appeared to facilitate iterative improvement in model development, as P9 described: "When I have this tool, I can visualize the results and see if the model meets the (fairness) need or not," and P10 confirmed: "The visualization gives convenience for me to understand how biased my model is." These insights indicate that future behavior change interventions could further explore visual representations of fairness metrics and bias indicators, as visualization appears to be a powerful tool for making abstract fairness concepts more concrete and actionable in practitioners' workflows.

5.4 Theme III: Expand from the Focus of Traditional Metrics

Our analysis revealed a significant shift in how participants approached model evaluation, moving beyond traditional performance metrics to incorporate fairness considerations. P7 acknowledged this transformation, stating "In data science practices, we mainly rely on the traditional metrics (accuracy) to evaluate the model, and tend to ignore the fairness issues." P8 echoed this sentiment, noting "Without Aequitas, my approach to this problem would be quite different. Normally I would not take fairness into my consideration, I only care about accuracy and F-1." This evolution in thinking was also reflected in P9's comment: "Previously, I just tried different models and (saw) if they are good (in terms of accuracy) and then delivered to the customers." This theme reinforces the goal that behavior change interventions can be designed to explicitly challenge and expand traditional evaluation frameworks, by integrating fairness metrics alongside conventional performance metrics in standard evaluation dashboards and workflows.

5.5 Theme IV: Integrate into Existing DS Workflows

Participants consistently expressed strong intentions to incorporate fairness considerations into their future work, suggesting the potential for longer-term impacts of the interventions. P8 stated, "I want to apply Aequitas in my future ML practices especially within contexts that come with real-world impact," while P10 similarly affirmed, "I will use this tool in my future projects." The perceived ease of integration appeared to be a crucial factor in these intentions, as P7 observed, "Aequitas seems very easy to be adapted to different usage scenarios; I would be very willing to incorporate it to my fairness-sensitive usage cases in the future." P11 specifically highlighted potential applications: "I think this tool should be used by banks when they develop loan prediction or fraud detection models." These findings highlight the importance of designing behavior change interventions that seamlessly integrate into existing workflows and tools. It suggests that future interventions should prioritize compatibility with common data science platforms and frameworks to facilitate long-term adoption.

5.6 Theme V: Boost Intrinsic Motivation

The interventions appeared to significantly influence participants' motivation to invest additional effort in ensuring model fairness. P1 noted, "It definitely motivates me to explore more about what's going on with this bias," and elaborated that this motivation led to concrete actions: "Potentially that (reading the Prime) is why I used multiple models and dove deeper into one model to conduct parameter tuning to refine the results." P2 similarly reported, "I definitely put more effort within this study to ensure fairness (after reading the Prime)." This increased motivation often translated into more comprehensive model development approaches, as P6 described: "These efforts include trying out different models and searching for tunable hyperparameters. Additionally, by tuning some hyperparameters, I was able to reduce the bias of the model but with a little sacrifice of model performance." One thing we can learn from this theme is that behavior change interventions should be designed not just to provide tools or guidelines, but to also consider actively cultivating intrinsic motivation by demonstrating the meaningful impact and professional value of responsible practices.

These qualitative findings complement our quantitative results by illuminating the psychological and practical mechanisms through which behavior change interventions influence data science practices. They suggest that effective interventions should consider personal connection, visual feedback, evolution of evaluation criteria, and motivation as key elements in promoting responsible data science practices. Furthermore, these findings highlight the importance of supporting data scientists in navigating the complex trade-offs inherent in responsible data science practice while maintaining their engagement and motivation.

6 Discussion and Future Work

Our findings provide valuable insights into the efficacy of behavior change interventions (BCIs) in promoting responsible data science (RDS) practices. By evaluating Prime (motivational priming) and Aequitas (fairness toolkit), we identified distinct ways in which these interventions influence fairness-oriented decision-making and practitioner behavior. This section discusses the broader implications of these findings, the trade-offs involved, and directions for future research.

Designing More Relatable Interventions: Our study highlights the role that motivation can play in fostering responsible data science practices. The Prime intervention, which framed fairness as a tangible and urgent issue, significantly influenced participants' motivation to adopt responsible behaviors. Interestingly, three female participants in the Prime group explicitly mentioned that they could relate to the disadvantaged groups described in the task. P3 reflected: *"I can totally relate to the situation [female applicants unfairly treated by loan approval models] as a female."* This suggests that interventions leveraging lived experiences or empathy may be particularly effective in motivating ethical decision-making, especially when practitioners identify with the affected groups. However, while motivational priming raised awareness, its impact on fairness metrics **(H3)** was not statistically significant. This underscores a key challenge: motivation alone may not suffice to translate ethical intentions into actionable outcomes without complementary tools or guidance.

Therefore, we need to move beyond motivation-only approaches toward comprehensive intervention strategies that address the full spectrum of behavioral change factors. Our findings suggest several directions for strengthening and complementing motivational interventions in the broader RDS field. First, *hybrid interventions* that combine motivational framing with embedded technical capabilities could optimize both engagement and effectiveness—for example, integrating empathy-building scenarios directly within fairness auditing tools. Second, *organizational scaffolding* is crucial, as individual motivation must be supported by institutional incentives, performance metrics that value fairness outcomes, and dedicated time allocation for responsible practices. Finally, the RDS field needs to explore *adaptive intervention systems* that can personalize motivational content based on practitioners' backgrounds, experiences, and the specific contexts in which they work. These directions acknowledge that sustainable behavior change requires not just individual motivation, but systemic support structures that make responsible practices both personally meaningful and professionally viable.

Designing Cognitively Efficient Interventions: The Aequitas intervention demonstrated superior results in promoting responsible behaviors **(H1)** and improving fairness metrics **(H3)**, but it also introduced a higher cognitive load **(H5)**. Participants reported increased mental demand and effort when using the toolkit, as it required additional steps for bias auditing and fairness corrections. For example, P7 remarked, *"It [Aequitas] is like a forcing function to let me revisit my model development to check and refine my model deployment."* Rather than viewing this trade-off as inevitable, we propose two concrete strategies for designing cognitively efficient interventions that maintain effectiveness while balancing cognitive load. *Progressive fairness workflows* could break complex auditing into smaller, manageable steps integrated with existing model development phases. Instead of a comprehensive post-hoc analysis, the tool could prompt simple fairness checks at natural breakpoints: data exploration ("Check for representation gaps"), feature engineering ("Identify potentially problematic correlations"), and model evaluation ("Assess prediction disparities"). Each step would require minimal additional effort while building toward a comprehensive fairness assessment. Secondly, *collaborative fairness dashboards* could distribute cognitive load across team members, allowing domain experts to focus on bias interpretation while technical practitioners handle implementation. These design principles acknowledge that sustainable adoption requires tools that enhance rather than complicate existing data science workflows, making responsible practices feel like natural extensions of balanced technical practice rather than additional burdens.

Weighing Costs and Benefits: Our results suggest that more demanding interventions like Aequitas may be warranted in high-stakes decision contexts (healthcare, lending, criminal justice), when fairness outcomes significantly impact vulnerable populations, or when organizational incentives explicitly value equitable practices. The cognitive burden becomes more acceptable when practitioners personally connect with fairness concerns—as demonstrated by participants who identified with disadvantaged groups. However, this tradeoff may be optimized through strategic application: employing high-effort interventions during critical development phases while using lightweight approaches for routine workflows. Future intervention designs could address this tension by automating repetitive fairness checks while preserving meaningful human judgment for complex ethical decisions.

Ecological Validity: This study employs a controlled experimental design to isolate the causal effects of specific interventions on practitioner behavior—a necessary first step in understanding behavior change mechanisms. While this approach enables rigorous comparison of intervention types, it necessarily involves tradeoffs with ecological validity: participants worked on well-defined tasks without the competing organizational pressures, career incentives, and time constraints that characterize real-world data science practice. Controlled studies allow us to establish whether interventions can influence behavior under ideal conditions, providing an upper bound on effectiveness and revealing mechanisms that may be obscured in field settings where multiple confounding factors operate simultaneously. Our findings demonstrate the potential efficacy of these intervention approaches, with the limitation that effect sizes may be attenuated in practice where fairness concerns must compete with other priorities. Future work should complement these controlled findings with field studies examining intervention adoption in authentic organizational contexts.

Future Work: Future research should extend beyond our current findings on intervention efficacy evaluation. Future work should explore longitudinal studies to assess the sustainability of behavior change interventions outside controlled environments. Key questions include: (1). How do workplace culture and time constraints affect the long-term adoption of tools like Aequitas? (2). Can motivational priming remain effective when ethical considerations compete with other priorities, such as model performance or deadlines? Additionally, investigating hybrid interventions—combining motivational framing with lightweight, embedded tooling—could optimize both motivation and usability. For example, integrating fairness alerts into existing data science platforms (e.g., Jupyter notebooks) might reduce cognitive load while maintaining ethical engagement. Lastly, expanding the scope of BCIs to include organizational incentives (e.g., tying fairness metrics to performance evaluations) could address systemic barriers identified in prior work[21].

Furthermore, an important consideration when interpreting our cognitive load findings is the distinction between different types of cognitive burden. The increased mental demand and effort observed with Aequitas could stem from two separate sources: (1) the inherent complexity of grappling with fairness concepts in data science work, or (2) the specific interface and workflow demands of the Aequitas tool itself. Cognitive load theory distinguishes between intrinsic cognitive load (essential to the task), extraneous cognitive load (imposed by the instructional design), and germane cognitive load (related to schema construction) [35]. Future work should aim to disentangle these factors to determine whether the observed load increase represents necessary engagement with fairness concepts (intrinsic/germane) or tool-specific complexity that could be optimized (extraneous). Such distinctions would help develop interventions that maximize meaningful cognitive engagement with fairness while minimizing unnecessary workflow friction.

7 Limitations

Our study offers valuable insights into behavior change interventions for responsible data science, but several limitations should be acknowledged. First, we opted for a 4 point likert scale for H2 (ranging from -2 to 2, without a neutral option) which may have confused participants who perceived a neutral impact and thus compromised the reliability of the COM-B factor measurement **(H2)**. Furthermore, To assess **H5**, cognitive load was measured using the NASA-TLX on a 7-point scale (0 = low demand to 7 = high demand) rather than the standard 20-point scale. Due to this non-standard scaling approach, we standardized responses before analysis and focused on within-subjects comparative analysis rather than absolute values, which allows for valid internal comparisons while potentially limiting direct comparison with studies using the traditional scale. Second, our sample size of 12 data scientists, while providing rich qualitative insights, limits the statistical power of our analysis. Future work should scale these evaluations with larger, more diverse participant pools across different organizational contexts. Third, the controlled laboratory setting of our experiment may not fully capture the complexities of real-world data science workflows, where organizational priorities, time constraints, and collaborative dynamics influence decision-making. The ecological validity of our findings would be strengthened through longitudinal field studies examining intervention adoption in authentic workplace environments. Fourth, our evaluation focused on two specific datasets (German Credit and Census Income) which may not represent the full spectrum of fairness challenges encountered in practice. Different domains and data types might introduce unique considerations that our current interventions do not address. Moreover, our within-subjects design may introduce potential learning effects that operate through multiple mechanisms. While we counterbalanced which dataset participants encountered first (Credit vs. Census), all participants completed the Control condition in their first session and the intervention condition in their second session. Participants' second session (intervention condition) may have benefited from practice with the study format despite the 24-hour separation and different datasets. This learning effect encompasses both general procedural familiarity and potential cross-task transfer between datasets, regardless of which specific intervention they received in their second session. This may cause our cognitive load findings (H5) to underestimate the true cognitive demands of interventions, particularly for Aequitas, as participants may have developed more efficient workflows by their second session. Future studies should employ fully counterbalanced designs that control for both dataset order and intervention order to explicitly measure and separate these learning effects. Finally, we evaluated behavioral changes and outcome improvements in a single session, which cannot capture the long-term sustainability of these effects. Future research should employ fully counterbalanced designs to examine whether the observed behavior changes persist over time and how they evolve as practitioners gain familiarity with interventions like Aequitas. While our upper-bound framing combined with limited sample size results in wide confidence intervals, this study establishes proof of concept that behavior change interventions can influence RDS practices—a necessary foundation for larger-scale studies. The large effect sizes observed suggest effects substantial enough to warrant further investigation despite interval width. We view this work as foundational evidence justifying investment in larger validation studies rather than definitive claims about real-world effectiveness.

8 Conclusion

Despite growing awareness of algorithmic harms, responsible data science practices remain inadequately adopted in practice. This study systematically evaluates behavior change interventions to bridge the gap between ethical principles and practitioner behaviors. In this paper, we conducted a study with 12 data scientists. We evaluated two behavior change interventions (BCIs)—Prime (motivational priming) and Aequitas (fairness toolkit)—to bridge the gap between ethical principles and responsible practice. We found that both interventions increased responsible behaviors, with Aequitas significantly improving fairness metrics (though at higher cognitive load) and Prime boosting motivation without compromising accuracy and cognitive load. The interventions were particularly effective when participants could personally relate to the fairness scenarios, suggesting that empathy plays a role in ethical decision-making. Importantly, our findings challenge the common assumption that fairness necessarily comes at the cost of model performance. These results highlight the need for balanced BCIs that combine technical tooling with motivational support to foster sustainable responsible data science. Future work should explore hybrid approaches that mitigate cognitive load while maintaining ethical engagement, as well as longitudinal adoption in real-world workflows where competing priorities exist. Critically, the field needs comprehensive intervention ecosystems beyond tools like Aequitas to address the entire data science pipeline, ensuring responsible practices become embedded across all phases of model development.

Acknowledgments

This work was supported by the National Science Foundation under grants IIS-2340539, IIS-2141506, IIS-2402718.

References

[1] Cecilia Aragon, Shion Guha, Marina Kogan, Michael Muller, and Gina Neff. 2022. *Human-centered data science: an introduction.* MIT Press.

[2] Agathe Balayn, Mireia Yurrita, Jie Yang, and Ujwal Gadiraju. 2023. "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.* 482–495.

[3] Barry Becker and Ronny Kohavi. 1996. Adult. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5XW20.

[4] Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilović, et al. 2019. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development* 63, 4/5 (2019), 4–1.

[5] Gema Bello-Orgaz, Jason J Jung, and David Camacho. 2016. Social big data: Recent achievements and new challenges. *Information Fusion* 28 (2016), 45–59.

[6] Sarah Bird, Miro Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, and Kathleen Walker. 2020. Fairlearn: A toolkit for assessing and improving fairness in AI. (2020).

[7] Longbing Cao. 2022. Ai in finance: challenges, techniques, and opportunities. *ACM Computing Surveys (CSUR)* 55, 3 (2022), 1–38.

[8] Leo Anthony Celi, Jacqueline Cellini, Marie-Laure Charpignon, Edward Christopher Dee, Franck Dernoncourt, Rene Eber, William Greig Mitchell, Lama Moukheiber, Julian Schirmer, Julia Situ, et al. 2022. Sources of bias in artificial intelligence that perpetuate healthcare disparities—A global review. *PLOS Digital Health* 1, 3 (2022), e0000022.

[9] Jiahao Chen, Victor Storchan, and Eren Kurshan. 2021. Beyond fairness metrics: Roadblocks and challenges for ethical ai in practice. *arXiv preprint arXiv:2108.06217* (2021).

[10] Anamaria Crisan, Brittany Fiore-Gartland, and Melanie Tory. 2020. Passing the data baton: A retrospective analysis on data science work and workers. *IEEE Transactions on Visualization and Computer Graphics* 27, 2 (2020), 1860–1870.

[11] André F Cruz, Catarina Belém, Sérgio Jesus, João Bravo, Pedro Saleiro, and Pedro Bizarro. 2022. Fairgbm: Gradient boosting with fairness constraints. *arXiv preprint arXiv:2209.07850* (2022).

[12] Wesley Hanwen Deng, Manish Nagireddy, Michelle Seng Ah Lee, Jatinder Singh, Zhiwei Steven Wu, Kenneth Holstein, and Haiyi Zhu. 2022. Exploring how machine learning practitioners (try to) use fairness toolkits. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency.* 473–484.

[13] Catherine D'ignazio and Lauren F Klein. 2023. *Data feminism.* MIT press.

[14] Ziwei Dong. 2025. *Towards Responsible Data Science with Behavior Change Interventions.* Ph. D. Dissertation.

[15] Ziwei Dong, Teanna Barrett, Ameya Patil, Yuichi Shoda, Leilani Battle, and Emily Wall. To appear 2025. A Design Space of Behavior Change Interventions for Responsible Data Science. *ACM Conference on Intelligent User Interfaces (IUI)* (To appear 2025).

[16] Ziwei Dong, Ameya Patil, Yuichi Shoda, Leilani Battle, and Emily Wall. To appear 2025. Behavior Matters: An Alternative Perspective on Promoting Responsible Data Science. *ACM SIGCHI Conference on Computer-Supported Cooperative Work & Social Computing (CSCW)* (To appear 2025).

[17] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining.* 259–268.

[18] Avijit Hazra. 2017. Using the confidence interval confidently. *Journal of Thoracic Disease* 9 (10 2017), 4124–4129. https://doi.org/10.21037/jtd.2017.09.14

[19] Andrés Domínguez Hernández and Vassilis Galanos. 2022. A toolkit of dilemmas: Beyond debiasing and fairness formulas for responsible AI/ML. In *2022 IEEE International Symposium on Technology and Society (ISTAS)*, Vol. 1. IEEE, 1–4.

[20] Hans Hofmann. 1994. Statlog (German Credit Data). UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5NC77.

[21] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. 2019. Improving fairness in machine learning systems: What do industry practitioners need?. In *Proceedings of the 2019 CHI conference on human factors in computing systems.* 1–16.

[22] Sérgio Jesus, Pedro Saleiro, Beatriz M Jorge, Rita P Ribeiro, João Gama, Pedro Bizarro, Rayid Ghani, et al. 2024. Aequitas Flow: Streamlining Fair ML Experimentation. *arXiv preprint arXiv:2405.05809* (2024).

[23] Daniël Lakens. 2013. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Frontiers in psychology* 4 (2013), 863.

[24] Michelle Seng Ah Lee and Jat Singh. 2021. The landscape and gaps in open source fairness toolkits. In *Proceedings of the 2021 CHI conference on human factors in computing systems.* 1–13.

[25] Suyun Liu and Luis Nunes Vicente. 2022. Accuracy and fairness trade-offs in machine learning: A stochastic multi-objective approach. *Computational Management Science* 19, 3 (2022), 513–537.

[26] Brandon May, Marek Palace, Rebecca Milne, Gary Dalton, Amy Meenaghan, and Sylvia Terbeck. 2025. Virtue, choice, and storytelling: how ethics, decision modalities and narrative framing influence decision inertia in a 360 degree extended reality environment. *Cognition, Technology & Work* 27, 3 (2025), 611–633.

[27] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)* 54, 6 (2021), 1–35.

[28] Susan Michie, Maartje M Van Stralen, and Robert West. 2011. The behaviour change wheel: a new method for characterising and designing behaviour change interventions. *Implementation science* 6 (2011), 1–12.

[29] Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. 2021. Algorithmic fairness: Choices, assumptions, and definitions. *Annual review of statistics and its application* 8, 1 (2021), 141–163.

[30] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. 2017. On fairness and calibration. *Advances in neural information processing systems* 30 (2017).

[31] Erasmo Purificato, Flavio Lorenzo, Francesca Fallucchi, and Ernesto William De Luca. 2023. The use of responsible artificial intelligence techniques in the context of loan approval processes. *International Journal of Human–Computer Interaction* 39, 7 (2023), 1543–1562.

[32] Christopher Rigano. 2019. Using artificial intelligence to address criminal justice needs. *National Institute of Justice Journal* 280, 1-10 (2019), 17.

[33] L Nelson Sanchez-Pinto, Yuan Luo, and Matthew M Churpek. 2018. Big data and data science in critical care. *Chest* 154, 5 (2018), 1239–1248.

[34] Hong Shen, Wesley H Deng, Aditi Chattopadhyay, Zhiwei Steven Wu, Xu Wang, and Haiyi Zhu. 2021. Value cards: An educational toolkit for teaching social impacts of machine learning through deliberation. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency.* 850–861.

[35] John Sweller. 2011. Cognitive load theory. In *Psychology of learning and motivation.* Vol. 55. Elsevier, 37–76.

[36] Gareth Terry, Nikki Hayfield, Victoria Clarke, Virginia Braun, et al. 2017. Thematic analysis. *The SAGE handbook of qualitative research in psychology* 2, 17-37 (2017), 25.

[37] Vladimira Timkova, Daniela Minarikova, Lubomira Fabryova, Jana Buckova, Peter Minarik, Zuzana Katreniakova, and Iveta Nagyova. 2024. Facilitators and barriers to behavior change in overweight and obesity management using the COM-B model. *Frontiers in Psychology* 15 (2024), 1280071.

[38] Sahil Verma and Julia Rubin. 2018. Fairness definitions explained. In *Proceedings of the international workshop on software fairness.* 1–7.

[39] Gianmario Voria, Stefano Lambiase, Maria Concetta Schiavone, Gemma Catolino, and Fabio Palomba. 2025. From expectation to habit: Why do software practitioners adopt fairness toolkits?. In *2025 IEEE/ACM 47th International Conference on Software Engineering: Software Engineering in Society (ICSE-SEIS).* IEEE, 94–105.

[40] Zijie J Wang, David Munechika, Seongmin Lee, and Duen Horng Chau. 2024. SuperNOVA: Design Strategies and Opportunities for Interactive Visualization in Computational Notebooks. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems.* 1–17.

[41] Iyad Zayour and Timothy C Lethbridge. 2001. Adoption of reverse engineering tools: a cognitive perspective and methodology. In *Proceedings 9th International Workshop on Program Comprehension. IWPC 2001.* IEEE, 245–255.

[42] Ellen Zegura, Carl DiSalvo, and Amanda Meng. 2018. Care and the practice of data science for social good. In *Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies.* 1–9.