Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Ziwei Dong Date

Towards Responsible Data Science with Behavior Change Interventions

By

Ziwei Dong
Doctor of Philosophy

Computer Science and Informatics

Emily Wall, Ph.D.
Advisor

Leilani Battle, Ph.D.
Committee Member

Joyce Ho, Ph.D.
Committee Member

Chinmay Kulkarni, Ph.D.
Committee Member

Accepted:

Kimberly Jacob Arriola, Ph.D., MPH
Dean of the James T. Laney School of Graduate Studies

April 25, 2025

Date

Towards Responsible Data Science with Behavior Change Interventions

By

Ziwei Dong
B.E., Nanjing University of Science and Technology, Nanjing, China, 2016
M.Sc., Northwestern University, IL, USA, 2019

Advisor: Emily Wall, Ph.D.

An abstract of
A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Computer Science and Informatics
2025

Abstract

Towards Responsible Data Science with Behavior Change Interventions
By Ziwei Dong

Data science holds immense potential for societal progress; yet, if unchecked, it can perpetuate harm and inequity. We have witnessed instances where data-driven systems mislabel individuals, leading to dehumanization and unequal access to essential resources. While the academic community has made strides in addressing these issues, the predominant focus has been on technical solutions around algorithmic fairness, often overlooking the people and systems involved. This thesis presents a novel approach to bridging this gap. It introduces three key elements: (1) the translational application of behavior change theories for promoting responsible data science practices, (2) a design space to scaffold the development of behavior change interventions in the data science context, and (3) the implementation and empirical assessment of behavior change interventions designed to meet the specific demands of responsible data science. This work extends beyond technical solutions to address the systemic issues at the core of responsible data science, presenting a series of works that ensures data science serves society responsibly.

Towards Responsible Data Science with Behavior Change Interventions

By

Ziwei Dong
Computer Science and Informatics
Emory University

Advisor: Emily Wall, Ph.D.

A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Computer Science and Informatics
2025

ACKNOWLEDGEMENTS

First and foremost, I wish to express my profound gratitude to my advisor, Emily, whose guidance, mentorship, understanding, motivation, inspiration, and unwavering support have been the cornerstone of my academic journey. The growth, confidence, and independence I have gained throughout this Ph.D. journey constitute a lifelong treasure that has already transformed me and will continue to enrich my life in the years to come.

I am sincerely grateful to my committee members, Leilani, Chinmay, and Joyce, whose critical feedback strengthened and shaped my research over the last few years. Your insights and guidance were invaluable to this work. My heartfelt appreciation extends to my collaborators and labmates—Ameya, Teanna, Eli, Caroline, Yuichi, Keke, Yanan, Shiyao, Thomas, and Mengyu—with whom I have had the privilege of brainstorming ideas and working closely. The intellectual discourse, collaborative spirit, and camaraderie we shared have made this journey both intellectually stimulating and personally rewarding.

To Olivia, your support and companionship throughout my Ph.D. have been an indispensable anchor in this challenging conquest. I am eternally grateful to my parents, Yang and Xiqin, whose unconditional love and unwavering belief in my abilities have provided the foundation upon which I could pursue my academic aspirations. My cousin Lu and cousin-in-law Hao, your emotional support and steadfast care throughout this demanding process have been a source of strength. I also extend my gratitude to other family members who have shown their love and encouragement: my grandparents Jufang and Minghu, my aunt and uncle Huiqin, Suogeng, Yiping and Jun, and my niece Lexi.

I am grateful for the serendipitous connections that brought extraordinary mentors and friends into my life: Zhiyuan, Abhishek, Jing, Rongmei, Pengfei, Jinfei, Chen, Shibo, Alejandro, and Shaojun. Thank you to all mentors who taught, trained,

enlightened, and shaped me during this journey. I am grateful for how you embodied the qualities I aspire to cultivate and the person I strive to become.

I wish to acknowledge the failures, rejections, frustrations, anxieties, perplexities, uncertainties, and sleepless nights that built my resilience to setbacks. These challenges illuminated the reality that life's path is rarely smooth. They honed my determination, strengthened my spirit, and cultivated capabilities I never knew I possessed. While knowledge may evolve with time, this internal growth remains an enduring gift of this journey.

Lastly, I extend gratitude to myself—for the unwavering dedication through countless late nights, for persisting when results seemed elusive, and for finding courage when doubt crept in. I thank myself for maintaining intellectual curiosity when fatigue set in, for adapting to setbacks with resilience, and for preserving the passion that initiated this journey. Many years later, Ziwei, if you look back on this thesis, I hope you can always be proud of the scholar—and the person—you became through this transformative experience.

Contents

List of Figures

List of Tables

List of Algorithms

Chapter 1

Introduction

1.1 Motivation

While data science can advance important societal goals, such as fighting climate change and species extinction, it can also cause considerable societal harm [16]. Individual mispredictions can lead to the dehumanization of Black people by labeling them as gorillas [135], or loss of health benefits for those who need them the most [61]. These examples hint towards larger systems of inequity that data science pipelines inadvertently perpetuate when left unchecked.

I have seen a heartening surge in academic research to counteract these inequities in machine learning and broader data science practices [123], including the introduction of the Conference on Fairness, Accountability, and Transparency in 2018, and numerous workshops such as Algorithmic Fairness through the Lens of Time at NeurIPS 2022 and the International Workshop on Responsible AI and Data Ethics (RAIDE) at the IEEE International Conference on Big Data 2022. Across these venues and others, existing work often focuses on ensuring that data science pipelines, and consequently their outputs, are mathematically and statistically sound. Issues of bias and inequity are then framed as mitigating the erosion of technical quality, such as detecting and

counteracting biased input data or biased algorithms; for example, developing bias mitigation strategies to counter bias in face detection datasets [194, 30].

However, modifying the algorithms and models that data scientists use is not enough to solve such a systemic problem. I liken this solution to modifying cigarettes to prevent lung cancer rather than helping smokers quit smoking. A technical solution may be satisfactory for avoiding traditional cigarettes, but it does not help people avoid addictive behaviors. Similarly, while I observe that technical solutions are essential to successful data science, I also argue that they are *insufficient* for ensuring responsible outcomes in human-AI interactions. Biases appear within datasets and algorithms because *people* inadvertently put them there. When I focus on the *inputs* (data, algorithms) and *outputs* (inferences) and not on the *agents* involved (people and systems), I may miss the opportunity to more meaningfully address the underlying causes of the problems I seek to fix. The conceptual diagram in Figure 1.1 illustrates the assertion that both human factors and technical factors mutually contribute to responsible data science.

Towards the goal of advancing responsible data science, there are several possible pathways. Responsible data science education establishes essential ethical foundations, but primarily reaches practitioners before they enter the field. In-the-moment behavior change interventions, on the other hand, target decision points *within actual workflows*, potentially redirecting choices toward more responsible outcomes. Long-term habit development, meanwhile, seeks to transform episodic responsible decisions into ingrained practice. While these approaches are complementary, this thesis focuses on behavior change interventions. I argue that behavior change interventions represent a particularly promising research direction for promoting responsible data science for two key reasons. First, since model outcomes are direct products of their development processes, targeting practitioner practices through behavior change interventions can more effectively ensure responsible models than ed-

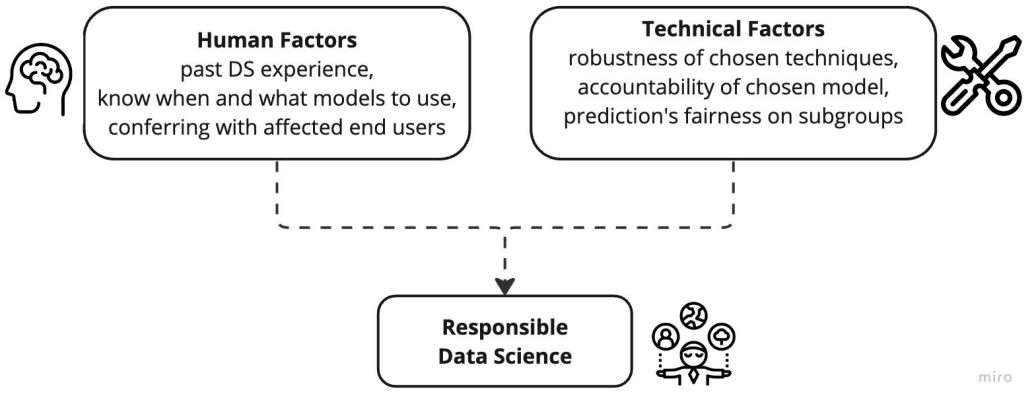

Figure 1.1: A conceptual diagram depicting the interdependent factors in responsible data science. This diagram depicts the complementary relationship between 'Technical Factors'—encompassing models, algorithms, and data science techniques—and 'Human Factors', which include past experience, decision making, and data processing.

ucational approaches alone. Second, interventions at decision time offer substantial leverage—they can influence critical choices precisely when practitioners face ethical dilemmas, complementing educational programs and potentially supporting and reinforcing the prolonged engagement necessary for habit formation. For these reasons, I posit that behavior change interventions are a compelling approach for improve responsible practice in real-world settings, particularly when considering the tension between outcome-focused approaches (which evaluate only final results through metrics like statistical parity) and process-oriented perspectives (which recognize that ethical development practices matter independently of outcomes). The behavior change framework I propose acknowledges that responsible data science warrants both fair outcomes and ethical processes, with practitioner behaviors serving as the critical link between the two.

To advance this high-potential research area of exploring in-the-moment behavior change interventions for promoting responsible data science, this thesis introduces (1) the translational application of behavior change theories tailored for promoting responsible data science practices by emphasizing the role of the behaviors and human

Thesis contribution: This thesis advances responsible data science by developing relevant theories, designing, developing and evaluating behavior change interventions that target the human factors behind data science decisions.

I: Develop Theories

RQ1: How can I utilize behavior change theories to inform a novel framework for responsible data science using the lens of behavior change interventions?
Contribution for Project 1: Theoretical Framework **(CSCW 2025 [58])**.

II: Synthesize Design Space

RQ2: How can I scaffold the design and development of behavior change interventions for responsible data science?

Contribution for Project 2: Design Space **(IUI 2025 [56]).**

III: Develop and Evaluate Interventions

RQ3: How effective are behavior change interventions at improving responsible data science outcomes

Contribution for Project 3: Intervention Development **(In preparation for TVCG 2026 [57])**.
Contribution for Project 4: Intervention Evaluation **(In preparation for CHI 2026 [59]).**

Figure 1.2: An overview of the dissertation research.

factors (data scientists), (2) a design space for behavior change interventions within the data science context, and (3) the design and evaluation of behavior change interventions for promoting responsible data science. These three components form an integrated approach that advances the research of human factors in responsible data science at different levels: establishing theoretical foundations, providing actionable design frameworks, and demonstrating practical effectiveness through developing and evaluating an intervention. Focusing on these complementary aspects creates a pathway from theory to practice that specifically targets the human behaviors underlying responsible data science decisions. Through these contributions, this thesis advances responsible data science research.

1.2 Summary of Thesis Research

This research aims to promote responsible data science through behavior change interventions. This consists of three high-level goals, summarized in Figure 1.2:

1. integrating theories from cognitive and clinical psychology with data science workflow knowledge and ethics guidelines to produce a **conceptual framework of responsible data science through the lens of behavior change theories**,
2. synthesizing a **design space** to provide guidance on how to design and develop behavior change interventions in the data science context, and
3. **developing and evaluating** the efficacy of behavior change interventions in the data science context to promote responsible data science.

This dissertation addresses these three goals across four projects. Specifically, I developed theories for behavior change interventions (**project 1**), synthesized a design space of behavior change strategies that can scaffold the design and implementation of future interventions (**project 2**), created a proof-of-concept visual analytic tool to facilitate responsible data science practices (**project 3**), and conducted a controlled evaluation of the efficacy of behavior change interventions (**project 4**).

1.3 Thesis Statement

In the pursuit of Responsible Data Science with Behavior Change Interventions, this thesis addresses the critical need to go beyond technical solutions. I aim to advance the field of responsible data science from the perspective of human behavior through three goals: (i) introducing behavior change theories tailored to promote responsible data science practices, (ii) establishing a comprehensive design space for behavior change interventions within the data science context, and (iii) developing

and evaluating specialized tools for behavior change aligned with the demands of responsible data science. By emphasizing the role of human agents and their behaviors in data science processes, I seek to address the systemic issues of bias and inequity that persist in AI and data science, going beyond mere technical solutions to foster meaningful and effective change in the field in the hands of responsible agents – data scientists. **This thesis advances responsible data science by developing relevant theories, designing, developing and evaluating behavior change interventions that target the human factors behind data science decisions.**

1.4 Research Questions

This thesis will address the following research questions, which correspond to the three goals described in section 1.2:

RQ1 How can I utilize behavior change theories to inform a novel framework for responsible data science using the lens of behavior change interventions? (Goal I: Develop Theories)

RQ2 How can I scaffold the design and development of behavior change interventions for responsible data science? (Goal II: Synthesize Design Space)

RQ3 How effective are behavior change interventions at improving responsible data science outcomes? (Goal III: Develop and Evaluate Interventions)

Chapter 2

Related Work

2.1 Data Science Process and Framework

Data science deals with extracting high-level knowledge from low-level data. It requires inter-disciplinary knowledge of statistics, computer science, and domain knowledge pertaining to the dataset at hand [57, 46, 25, 64]. In this section, we describe existing surveys characterizing different high-level steps in the data analysis process, which we refer to as "the data science pipeline".

One established characterization of the data science pipeline is the KDD model (knowledge discovery in databases) by Fayyad [65]. It is comprised of five phases: Data selection, Preprocessing, Transformation, Mining, and Interpretation or Evaluation. The CRISP-DM model (Cross-Industry Standard Processes for Data Mining) [191] was proposed to standardize the data science process at the industry level. CRISP-DM has six phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. In his reflections on data science as a field [57], Donoho identifies six high-level activities in the data science process: **Data gathering**, **Preparation and exploration**, **Data representation and transformation**, **Computing with data**, **Data modeling**, and **Data visu-**

alization and presentation. Amershi *et al.* [9] formalizes the execution pipeline for data science, which has been a ubiquitous tool in data science for the past decade. They identify nine stages in the ML workflow - Model requirements, Data collection, Data cleaning, Data labeling, Feature engineering, Model training, Model evaluation, Model deployment, and Model monitoring. We observe that these frameworks define high-level stages to have a self-explanatory meaning, and although labeled differently across frameworks, stages having similar functionality can be identified.

We also note key commonalities among the frameworks, namely: these various granularities of the framework are not mutually exclusive and serve mainly as high-level structures to guide development for specific stages of the pipeline. Crisan *et al.* [46] echo this similarity and accordingly model the data science pipeline as four high-level stages that commonly appear across other frameworks: Preparation, Analysis, Deployment and Communication, with two complementary phases - Collaboration and Pedagogy.

In the forthcoming subsections, we review three stages of model development including Pre-Processing, In-Processing, and Post-Processing, particularly highlighting how researchers have considered the concept of model fairness within these stages.

2.1.1 Pre-processing

Pre-processing is a critical first step in building a data science model. Data quality, purity, and representation have a significant impact on model performance and fairness. Before building a machine learning model, it is often beneficial to apply some pre-processing operations (e.g., normalization, data transformation, data reduction) on the dataset to enhance the data quality to improve the model's performance and generalizability. There are a number of objective factors that can contribute to biases in models related to the feature space[17], such as: (i) uneven data distribution within these groups, (ii) their label distribution may be variable, (iii) some data may

be incorrectly labeled, (iv) the input data dimensionality may not fit the model, (v) the distribution of data may be too sparse for a specific model to learn some groups' representations, or furthermore, (vi) the data itself may be the result of biased social climates, all of which may warrant bias-mitigation processing before being fed into the model[32].

Various pre-processing operations can influence the data differently. For instance, data normalization[41] was first proposed in Edgar F. Codd's rational database. It scales down the scope of the dataset into a pre-defined range of values. This technique is important for some scale-sensitive data science models, such as K-nearest neighbors and neural networks. Feature transformation is a family of data pre-processing techniques that transfer the feature space of the original dataset into an alternative, more compact dimension space. Feature transformation operations aim to simplify the data presentation while preserving as much information as possible. One canonical and widely used technique is Principle Component Analysis (PCA)[94]. It approximates a high-dimensional dataset with a lower-dimensional linear subspace. By creating uncorrelated feature variables that maximize variance, improve interpretability, and at the same time, minimize information loss. Fordor[70] summarized multiple frequently used dimension reduction techniques.

2.1.2 In-processing

In the pipeline of data science, in-processing plays a crucial role by integrating fairness and ethical considerations directly into the model-building process. This approach, as opposed to pre- or post-processing, involves modifying the learning algorithms themselves to reduce bias and ensure fairness. Key techniques in this domain include the use of regularization methods[103, 50, 76] that penalize unfairness in models, adversarial training that aims to make models robust against biased data, and ensemble methods[20, 96] that combine predictions from multiple models to mitigate

individual biases. One notable example of debiasing during in-processing is the work by Zhang et al.[197], where they proposed an adversarial debiasing framework that learns to predict accurately while reducing discrimination. However, in-processing methods for debiasing are not without challenges; they often involve a delicate balance between maintaining model accuracy and reducing bias. Furthermore, there is no one-size-fits-all solution, as the appropriate technique largely depends on the specific context and type of bias present in the data. The future of debiasing during in-processing in data science appears promising, with burgeoning research focusing on developing more sophisticated algorithms that can seamlessly integrate fairness without significantly compromising on model performance.

2.1.3 Post-processing

Post-processing in data science emerges as a pivotal strategy to address fairness and ethical concerns after a model has made its predictions. This technique is particularly useful when altering the data or training process is not feasible or has been insufficient to address biases. Post-processing debiasing methods typically involve adjusting the decision thresholds of a trained model or recalibrating its outputs to ensure fair treatment across different groups[83, 96, 126, 172]. A notable instance is the work by Hardt et al.[83], who introduced an equality of opportunity model that adjusts thresholds for different demographic groups to achieve fairness in predictions. While these methods are effective in enhancing fairness in the short term, they come with their own set of challenges. For instance, indiscriminate application of post-processing debiasing can lead to a decrease in overall model accuracy and can sometimes be perceived as unfair by individuals belonging to the majority group[154]. Moreover, determining the appropriate adjustments often requires a deep understanding of both the model's behavior and the socio-cultural context of the data. Looking forward, debiasing during post-processing is an area poised for growth,

with research increasingly focusing on developing more nuanced and context-aware methods that can ensure fairness without significantly diminishing the utility of the model's predictions.

2.2 Responsible Data Science

Analyzing data and drawing insights from them places significant power and consequently responsibility in the hands of the analyst since these insights can potentially affect policies on a large scale. For example, when opaque models are used to make court rulings, it becomes impossible to understand why a decision was made or how biased inputs such as insufficient training data influenced the outcome [15, 7]. Having realized this responsibility, the data science community has called for more deliberation on the ethics of current data science pipelines. Van der Aalst [174] described that responsible data science centers around four challenging questions: (1) fairness: data science without prejudice - how to avoid unfair conclusions even if they are true, (2) accuracy: data science without guesswork - how to answer questions with a guaranteed level of accuracy, (3) confidentiality: data science that ensures confidentiality - how to answer questions without revealing secrets, and (4) transparency: data science that provides transparency - how to clarify answers so that they become indisputable? These questions characterize the fundamental challenges that responsible data science is facing.

Data science methods learn from training data with the aim of optimizing an objective, such as maximizing correct classifications. However, achieving this objective does not guarantee responsibility and accountability. Training data can carry biases, resulting in under-representation or discrimination against certain minority groups. Even when sensitive attributes are excluded, systematic rejection of specific groups may persist. Profiling can exacerbate the stigmatization of these groups.

Consequently, it is essential to develop approaches for identifying irresponsible and unethical decisions, like unintended discrimination, and devise methods to promote fairness. Ehsan *et al.*[60] argued that actionable interventions are in demand to recognize both the affordances and potential pitfalls of data science and AI. As a critical first step of developing behavior change interventions for responsible data science, we must first operationalize the concept. To this end, we further characterize the key ingredients of responsible data science in chapter 3.

2.3 Theories of Behavior Change

In this section, we illustrate how existing psychological models can be applied to inform the design of behavior change interventions in data science. There have been numerous applications of behavior change techniques in the space of personal health such as for smoking cessation [27, 141] and in environmental domains such as for managing carbon footprint [137, 148]. Among this class of theories, there are a few key concepts that are useful to operationally define before diving into the theories. A *target behavior* is the behavior that we seek (such as smoking cessation), which we achieve through *behavior change*. Behavior change *interventions* are designed to encourage or influence target behaviors.

In spite of the application of behavior change interventions in numerous domains, a survey by Wiafe *et al.* [190] revealed that only half of the behavior change interventions in persuasive systems across the domains of health, commerce, education, and environment have a theoretical grounding. Orji *et al.* [139] revealed a similar finding for persuasive technologies in the clinical domain. Furthermore, prior work suggests that behavior change interventions informed by psychological theories are more effective than those that are not [127, 33], promoting what is known as *evidence-based practices*. Accordingly, there has been substantial theoretical development on

evidence-based behavior change interventions. These theories of behavior change can be broadly classified into 3 categories:

1. **Factors Affecting Behavior Change** which tell us about the individual or group-level characteristics that can influence the likelihood of a target behavior being achieved,
2. **Behavior Change Techniques (BCT)** which are specific techniques or interventions that, leveraging particular factors, can increase the likelihood of a target behavior, and
3. **Mechanisms of Action** which explain the underlying cognitive mechanism that makes a specific factor or technique work to influence behavior.

That being said, many theories in each category have overlapping constructs [67, 127, 12, 33, 130, 3] which have been shown to make it difficult to identify individual processes or factors underlying successful behavior change [139]. Further, most of these theories are rooted in psychology and clinical research with limited empirically verified attempts at generalization across different fields. Thus, rather than comprehensively surveying theories, we instead focus on identifying and discussing the theories that appear most relevant for the data science context, as advised by Pinder *et al.* [143] and Michie *et al.* [127]. We accordingly choose theories that are highly cited and have more tangible implementations. In this thesis research, we exclude theories of behavior change from our review that are too specific to the application domain [66, 163, 187] and the theories designed primarily for longer-term intervention [68, 163].

We describe these categories of theories further in the next chapter.

Chapter 3

Developing Theories for Responsible Data Science through Behavioral Change Interventions

3.1 Motivation

Building on the motivation established in the chapter 1, this chapter delves deeper into the theoretical foundations of behavior change as applied to responsible data science. This chapter focuses on addressing a critical gap in current approaches to responsible data science.

There has been an encouraging increase in academic research focusing on addressing these inequities within machine learning and data science[123], yet these efforts remain primarily concentrated on the technical dimensions of the problem. Notable developments include the 2018 Conference on Fairness, Accountability, and Transparency and various workshops at significant conferences, but the human behavioral aspects often remain underexplored. While the technical solutions described in chapter 1 are necessary, they represent only half of the equation. We argue that

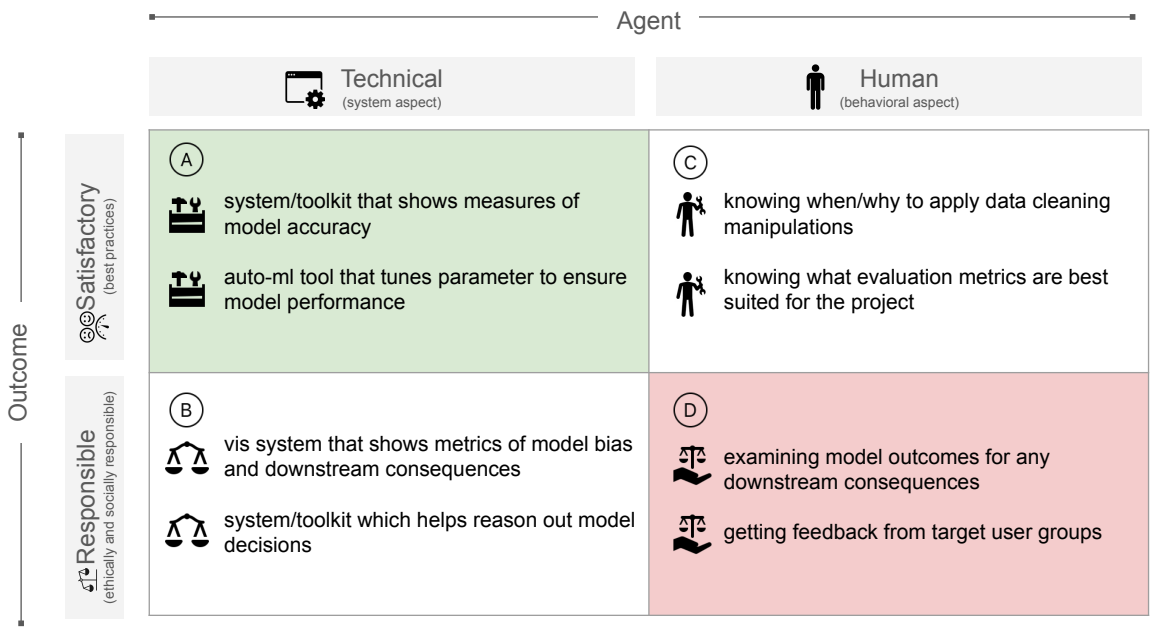

Figure 3.1: We characterize data science practices according to desired outcomes (rows – satisfactory and responsible) and agents (columns – technical and human). It is important to note that outcomes are not mutually exclusive. Rigorous data science has historically emphasized technical aspects like auto-tuning and measures of model accuracy (A, green cell). Recent efforts towards model fairness have illustrated responsible data science, but still ultimately rely on technical indicators and algorithmic solutions (B). In this project, we emphasize the agency of humans (C and D, right-hand column), and in particular, how human behaviors can contribute to responsible data science (D, red cell).

merely adjusting algorithms and models is insufficient to address the systemic nature of these issues. The spirit of this chapter is that meaningful progress in responsible data science requires attention to both technical systems and human behaviors, with particular emphasis on the underexplored responsible human behaviors in data science.

Hence in this project, we explore an alternative viewpoint in the literature. Specifically, **this project explores opportunities to redefine responsible data science to encompass not only technical responsibility (holding algorithms/ datasets accountable) but also behavioral responsibility, i.e., holding data scientists accountable for the *patterns of behavior* that may lead to positive or negative social outcomes**. To this end, we reframe existing literature on data science best practices and ethics guidelines through the lens of behavior change models from cognitive and clinical psychology. To ground our discussion, we draw parallels from successful behavior change interventions from cognitive and clinical psychology such as smoking cessation [27] to common data science scenarios today such as model training and exploratory visual data analysis. In this way, we show how several key principles from foundational behavior change research translate to the data science domain. This research project addresses **RQ 1:** *How can we utilize behavior change theories to inform a novel framework for responsible data science using the lens of behavior change interventions? (Goal I: Develop Theories)* This research project[55] has been accepted for publication at ACM SIGCHI CSCW 2025 as a full paper.

For reference, we include Table 3.1 below to summarize acronyms that will be introduced and used throughout the remainder of the project.

Table 3.1: A summary of the acronyms used throughout this project.

Acronym	Meaning
FBC	Factors affecting Behavior Change
BCT	Behavior Change Techniques
MoA	Mechanisms of Action
FBM	Fogg Behavior Model [72]
COM-B	Capability (C), Opportunity (O), and Motivation (M) - Behavior (B) [129]
TDF	Theoretical Domains Framework [127, 12]
BCTT	Behavior Change Techniques Taxonomy [3, 130]

3.2 Identifying Relevant Theories of Behavior Change for Data Science

In order to deliver responsible behaviors in data science, we seek to understand the heuristics behind effective behavior change techniques and transfer them into the data science domain. In this section, we illustrate how existing psychological models can be applied to inform the design of behavior change interventions in data science.

There have been numerous applications of behavior change techniques in the space of personal health such as for smoking cessation [27, 141], in environmental domains such as for managing carbon footprint [137, 148]. In spite of the application of behavior change interventions in numerous domains, a survey by Wiafe *et al.* [190] revealed that only half of the behavior change interventions in persuasive systems across the domains of health, commerce, education, and environment have a theoretical grounding. Orji *et al.* [139] revealed a similar finding for persuasive technologies in the clinical domain. Furthermore, prior works suggest that behavior change interventions informed by psychology theory are more effective than those that are not [127, 33], promoting what is known as *evidence-based practices*. Accordingly, there has been substantial theoretical development on evidence-based behavior change interventions.

To identify relevant behavior change theories, we conducted a literature search beginning with two canonical theories on factors influencing behavior change by Fogg

[72] and Michie et al. [129]. We collected relevant papers by searching forward- and back-references, as well as conducting additional keyword-based searches in Google Scholar, including keywords such as "behavior change theories" and "behavior change interventions." Throughout this exploratory process, we prioritized selecting theories and studies that are not only highly cited but have also stood the test of time (i.e., are still cited by a significant body of research at the time of this writing). From this corpus of relevant theories, we then grouped them into the following three categories:

1. **Factors Affecting Behavior Change (FBC)** which tell us about the individual or group-level characteristics that can influence the likelihood of a target behavior being achieved,
2. **Behavior Change Techniques (BCT)** which are specific techniques or interventions that, leveraging particular factors, can increase the likelihood of a target behavior, and
3. **Mechanisms of Action (MoA)** which explain the underlying cognitive mechanism that makes a specific factor or technique work to influence behavior.

That being said, many theories in each category have overlapping constructs [67, 127, 12, 33, 130, 3] which have been shown to make it difficult to identify individual processes or factors underlying successful behavior change [139]. Further, most of these theories are rooted in psychology and clinical research with limited empirically verified attempts at generalization across different fields. Thus, rather than comprehensively surveying theories of behavior change in this project, we instead focus on identifying and discussing the theories that appear most relevant for the data science context, as advised by Pinder *et al.* [143] and Michie *et al.* [127]. We accordingly choose theories that are highly cited and have more tangible implementations. In this section, we describe these theories and use them to characterize behavior change in the domain contexts listed in Figure 3.2 to ground them. We demonstrate how to use these theories to generate a series of interventions in section 3.5.

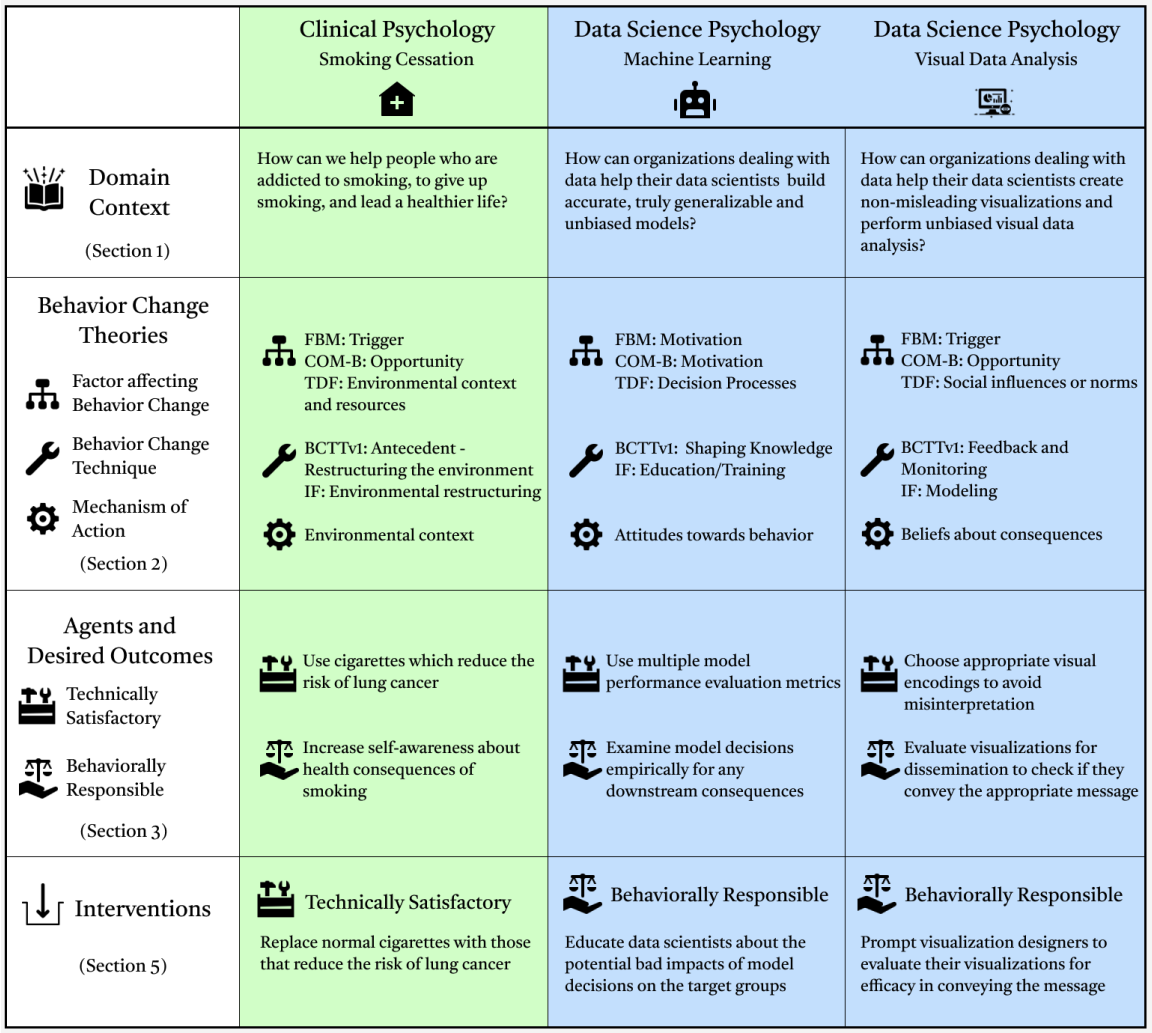

Figure 3.2: Drawing analogies from behavior change solutions in the clinical domain (green) to the data science domain (blue). Each column represents a behavior change domain. The rows characterize the behavior change problem and solutions, starting with the domain context. The next row characterizes exemplary theories of behavior change, followed by Agents and Desired Outcomes, and how together these might inform a specific intervention in each domain context (final row). The agents and outcomes, characterized as technically satisfactory or behaviorally responsible, are described further in Figure 5.1 and Section 3.3.1. We hand-pick these limited examples for the sake of space and to demonstrate how behavior change theory can be applied across different domains to bring about the desired outcome through the agent in a generalizable way.

3.2.1 Factors Affecting Behavior Change (FBC)

Here, we summarize established theories describing key factors that influence behavior change. While certainly not exhaustive, we focus on the following three prominent theories because they are well-established in the literature and complementary within the context of data science.

1. ***Fogg Behavior Model*** [72] or FBM (Fogg Behavior Model) proposes that behavior is comprised of three primary components: *motivation, ability*, and *trigger*. *Motivation* comprises both conscious and unconscious cognitive processes that guide and stimulate behavior. *Ability* refers to an individual's psychological and physical ability to engage in a particular activity. *Trigger* is a cue or a call to a particular activity. In the FBM, a *trigger* represents a tangible event that, under the appropriate circumstances, prompts an individual to change their behavior.

2. ***COM-B Model*** [129], standing for *Capability (C)*, *Opportunity (O)*, and *Motivation (M)* is a behavior change model that identifies these three key factors as influential in modifying behavior (B). Although *motivation* and *capability* align with the meanings of *motivation* and *ability* in FBM, COM-B introduces an additional element called *opportunity*. *Opportunity* encompasses external factors that enable or hinder the performance of a behavior.

3. ***Theoretical Domains Framework*** [127, 12] or TDF identifies 14 empirically verified domains that contain different factors which affect behavior change. TDF [12] consists of 84 factors organized into these 14 domains. The domains include *knowledge*, *skill*, *social role and identity*, *benefits about capability*, *optimism*, *belief about consequence*, *reinforcement*, *intentions*, *goals*, *memory/attention/decision process*, *environmental context and resources*, *social influence*, *emotion* and *behavior regulation*. TDF extends its scope to focus on

external social and environmental factors, providing a more fine-grained framework for identifying factors affecting behavior change.

We found that the domains discussed in TDF closely relate to the success of data science. However, the TDF, although good at identifying fine-grained factors affecting behavior change, is difficult to operationalize compared to the COM-B model, and thus has seen fewer direct applications [143]. The conciseness and usefulness of the COM-B model is corroborated by the fact that most prior theories [67, 127, 33, 12] including TDF, ultimately break down into components of the COM-B model. Finally, while COM-B balances specificity and generalizability, the Fogg Behavior Model [72] is one of the first theories to consider Behavior Change Techniques, discussed next. In summary, we refer to these three theories of factors affecting behavior change because of the complementary balance they provide in being concise (COM-B), specific (TDF), and operationalizable (FBM).

In Figure 3.2, the target behavior for the smoking cessation example could use an increase in the *opportunity* as per COM-B and TDF, to avoid lung cancer by using risk-free cigarettes, thus leading a healthier life. As per FBM, risk-free cigarettes provide a *Trigger* to bring about the change. Under TDF, such a behavior change intervention falls under the *environmental context and resources category* as it alters the resources available at hand. Note that while this solution can reduce toxicity, it does not address the underlying addiction. We discuss this further in Section 3.3.3.

On the other hand for the data science domain, the target behavior for the machine learning example calls for an increase in *motivation* as per FBM and COM-B to consider and empirically verify the greater impacts of the decisions made from their deployed ML models. As per TDF, this falls under changing the *decision processes* of the data scientist to include this verification step in their workflow. The target behavior for the visual data analysis example could be achieved by providing a *trigger* (as per FBM) to the visualization designer to incorporate an evaluation step in the

workflow before publishing the visualization. This provides an *opportunity* (as per COM-B) to the designer to verify if their visualizations conform to the *social norms* (as per TDF) of creating visualizations and are therefore effective in conveying the message.

3.2.2 Behavior Change Techniques (BCT)

Behavior change techniques put the aforementioned factors of behavior change to work, implementing the interventions which can bring about behavior change. Michie *et al.* [129] provide a coarse categorization of these techniques as *intervention functions* (IF). However, the most detailed taxonomy in this regard — the Behavior Change Techniques Taxonomy (BCTTv1), was created by Abraham & Michie *et al.* [3, 130], which lists 93 such techniques clustered into 16 categories. We use the BCTTv1 taxonomy for its descriptive power and Michie *et al.*'s [129] categorisation for its conciseness, when designing interventions, as described in our illustrative examples in section 3.5.

In Figure 3.2, for achieving the target behavior, one could *restructure the environment* by providing access to risk-free cigarettes in the smoking cessation example. In the data science domain, to achieve the target behaviors in the machine learning example, organizations could *educate/train* (as per intervention functions) their data scientists to identify possible negative impacts of the decisions of their deployed models on the target groups. This *shaping of their knowledge* (as per BCTTv1) can induce a change in their workflows to include empirical verification of downstream consequences of their model decisions. In terms of the visual data analysis example, the target behaviors of evaluating visualizations can be achieved through reminding designers to *compare* (as per BCTTv1) their visualizations to the commonly accepted visualization design norms or with visualization design *models* (as per intervention functions) so that viewers do not have difficulties in understanding the conveyed

message with the appropriate data context.

3.2.3 Mechanisms of Action (MoA)

Mechanisms of Action (MoA) represent the processes through which a BCT affects behavior. In other words, it explains *how* a factor of behavior change influences a certain technique to bring about the target change. Carey *et al.* [34] identified 26 different mechanisms of action and linked the behavior change techniques from the BCTTv1 taxonomy to these mechanisms (prompting or giving cues to the subject works by leveraging the *attention* and behavioral cueing mechanisms of human cognition, both of which affect the *capability* of the subject). We refer back to these Mechanisms of Action to understand the most effective means of designing interventions (section 3.5) with a theoretical grounding, thereby maximizing impact.

In Figure 3.2, the behavior change techniques of restructuring the environment for the target behavior in the smoking cessation example works through a change in the *environmental context* of the individuals. In the data science context (Jupyter Notebook), training the data scientists about the potential ill-effects of their model decisions on the target groups helps in changing the *attitudes towards their behaviors*. One potential behavior change technique – social comparison in the visual data analysis example works by influencing the visualization designers to adhere to socially accepted *subjective norms* of visualization design while incorporating the data context.

3.3 Responsible Data Science

As a precursor to translating behavior change theories to responsible data science, we must identify what constitutes responsible data science. **Responsible Data Science** includes efforts that address both technical and societal issues. We operationally

adopt the definition of responsible data science from Cheng et al[39] which says the objective of Responsible Data Science is to address the social expectations of generating shared value – enhancing both data science models' ability and benefits to society. This definition aligns with existing research examining "Ethical AI" and related topics[122, 169, 36, 181]. In subsection 3.3.1, we characterize responsible data science as a function of *agents* and *outcomes*, with a (typically implicit) role of behavior which can influence the outcomes. While we briefly review technically satisfactory practices in data science in subsection 3.3.2, our primary focus of this project, elaborated in subsection 3.3.3, is on the aspect of behavioral responsibility.

3.3.1 Characterizing Agents and Outcomes of Responsible Data Science

To scaffold our discussion of responsible data science, we find it useful to characterize it into two dimensions (as shown in Figure 5.1): *agents* and *desired outcomes*. The first dimension, agent, can be *technical* or *human*. Technical agents represent systems or techniques used in data science that have the potential to influence the rigor of data science practice through technical indicators, algorithms, systems, and toolkits that are incorporated into the data science project. Human agents, on the other hand, represent behavioral actions that affect the rigor of data science practice. We choose this terminology to emphasize the proactive role of agents within data science and AI. However, our framework is not limited to only AI/data science. It can also be extended into other automated interventions designed with different application scenarios.

The second dimension, the desired outcome, indicates the extent of attention to care and responsibility paid in the data science practice. We categorize desired outcomes loosely as *satisfactory* and *responsible*. These outcomes are not mutually exclusive and can overlap. Satisfactory outcomes focus on **maximizing benefits**

by following the established best practices without much regard to ethics; a loan approval model that maximizes the profit of banks but treats applicants who come from different genders unfairly. Responsible outcomes, on the other hand, aim to **minimize harm** and actively benefit society, incorporating ethical considerations throughout the data science process, a face recognition model that works well for humans from different ethnic groups. Moreover, being responsible itself can be seen as an **attitude** within the data science process, guiding actions and decisions with the intent of delivering responsible results. A responsible data science practice can, and should, encompass both technically satisfactory and behaviorally responsible actions.

Among the four combinations of these dimensions shown in Figure 5.1, we highlight the complementary importance of "Technically Satisfactory" and "Behaviorally Responsible" practices in in the frame of responsible data science. "Technically Satisfactory" practices (Figure 1A, green cell) have traditionally been the focus of data science practitioners to ensure that technical aspects of model development are sound, using appropriate tools, models, and metrics. However, they often lack consideration of ethical implications. In contrast, "Behaviorally Responsible" practices (Figure 1D, red cell) emphasize the ethical responsibilities of data scientists and the broader societal impacts of their actions. This focus on human behavior addresses the root causes of biases and ethical issues that technical solutions alone cannot resolve. in more detail in Sections 3.3.2 and 3.3.3 and connect them to the examples in Figure 3.2.

3.3.2 Technically Satisfactory Practices for Responsible Data Science

Every step in the data science pipeline presents opportunities for decisions that can significantly influence outcomes. In this subsection, we delve into the specifics of technically satisfactory practices, elucidating key aspects that demand adherence to best practices based on findings from a comprehensive survey on bias and fairness in

machine learning [122].

1. **Applying appropriate statistical tests:** After a research hypothesis is formulated, a suitable statistical test must be used to verify it. However, since domain experts may not be well-versed in statistics, the selection of appropriate statistical tests (one-way or two-way ANOVA) and parameters like significance level must be carefully considered[22, 99].
2. **Applying proper data science models:** The choice of model can significantly impact the quality of results and the ability to make meaningful predictions or decisions [51, 107]. Depending on the nature of the data, different models may be more appropriate. Moreover, model selection should consider factors such as scalability, computational resources, and interpretability. Regularization techniques and hyperparameter tuning further refine model performance. In some cases, ensemble methods or domain-specific models may be preferred.
3. **Applying suitable evaluation metrics:** Applying appropriate evaluation metrics is a pivotal aspect of ensuring a technically sound data science project. It is crucial to align the choice of evaluation metrics with the project's specific objectives [199]. Depending on whether the task involves classification, regression, or clustering, different metrics such as accuracy, precision, recall, F1-score, or Mean Absolute Error (MAE) should be carefully considered. Additionally, the presence of imbalanced data or unique business considerations may warrant the use of specialized metrics. Domain knowledge and collaboration with subject matter experts can further guide the selection of metrics that best reflect the real-world impact of the data science solution.
4. **Visualizing or communicating results:** Correll [42] calls for communicating the results of data analysis sessions with consideration for the data context and uncertainties, especially when using the medium of visualizations, which can abstract or trivialize the context provided by the data. One example is the

proposed use of fuzzy gradient plots instead of well-defined bar charts to better convey the uncertainty in the data [43].

Row 2 in Figure 3.2 illustrates technical, but ethically blind practices. For the smoking cessation problem, the example of using specialized cigarettes that prevent the risk of lung cancer makes use of technological advancements to achieve the desired satisfactory outcome. Extending the analogy to the data science domain, we could use multiple model accuracy metrics to gauge model performance (point 3 in the aforementioned bullet list). In the visual data analysis example, using appropriate empirically verified visual encodings for designing visualizations leads us to the satisfactory outcome of designing good visualizations (point 4 in the aforementioned bullet list).

3.3.3 Behaviorally Responsible Practices in Data Science

Several behaviorally responsible approaches complement existing technically satisfactory practices such Aragon *et al.*'s ethical principles in Human-Centered Data Science [11], Heise *et al.*'s primary ethical norms in computational research [88], and Zegura *et al.*'s calls for care and social good when practicing data science. Among these literature, we emphasize insights from *Human-Centered Data Science*[11] and the concept of *Care and the practice of data science for social good*[195]. These approaches complement one another by emphasizing the importance of responsible motivation when practicing data science and characterize actionable solutions to facilitate ethical practices in data science. While *Human-Centered Data Science* emphasizes high-level guidelines for practicing data science with care and rigor, *Care and the practice of data science for social good* delineates responsible practices at each stage of the data science pipeline, such as problem understanding and data preparation, which we describe in greater detail.

First, *Human-Centered Data Science*[11] provides a foundational resource that

facilitates a systematic approach to contemplating behaviorally responsible practices within the realm of data science. The book offers a comprehensive set of ethical guidelines that encourage a nuanced consideration of data science projects, ethics on defining the data science problem and ethical principles of training, validating, and testing data science models. The authors emphasize the key characteristic of responsible data science: "Our goal here is to make you aware that thinking critically and caring about your process and how it affects your results, as well as the people whose behavior is represented in your dataset, is needed every step of the way" [11]. The applicability of these ethical guidelines is notably well-suited to a wide range of situations where humans are, or ought to be, involved, such as loan approval and criminal recidivism predictions.

Second, the authors of *Care and the practice of data science for social good*[195] argued responsible practices are informed by a thoughtful examination of *how* research is done and in what *context* it is done. It argues responsible data science relies on an ethics approach rooted in practicality: ethics involves not only adhering to formal rules or their definitions but also observing actual behaviors. Ethics shouldn't be treated as a goal to optimize or "manipulate."

We assert that ethics requires a continuous process of reflection—considering potential risks, benefits, and harms. Yet, thoughtfulness alone does not prevent harm. The imperative to embrace responsible behaviors in data science emerges from the recognition that a standardized checklist is often insufficient across diverse scenarios encountered in the field [11]. Instead, behaviorally responsible data science practices demand that practitioners proactively cultivate and dynamically respond to their specific data science problem and context. Thus, a reflexive and adaptable stance is essential, acknowledging that the ethical considerations surrounding each data science project are nuanced and distinct. This diversity emphasizes the pivotal role of "**Care Ethics**[153, 11, 195]", a key concept to both *Human-Centered Data Science* and the

concept of *Care and the practice of data science for social good* as a foundational principle guiding behaviorally responsible data science.

Care Ethics encourages data practitioners to approach their work with a deep sense of empathy and conscientiousness[195, 125]. For example, Zegura et al[195] proposed an orientation to a caring mindset in the practice of data science that facilitates social good; Meng et al[125] highlighted the importance of applying the ethics of care in democracy within collaborative data work. This approach prompts practitioners to reflect on how their choices would impact individuals, communities, and society at large. The notion of Care Ethics introduces a transformative shift in perspective[26]. Encouraging data scientists to envisage their data science projects as endeavors involving their own family and loved ones cultivates a heightened sense of responsibility. Promoting Care Ethics not only enhances the behavioral responsibility of data science but also infuses the decision-making process with an intrinsic sense of accountability.

Inspired by the principles contained within care ethics[153, 11, 195], we review some actionable activities that align with behaviorally responsible data science. These practices are not exhaustive and should not be viewed as a checklist – instead, these serve only as inspirational examples to further ground the concept of behavioral responsibility in data science.

1. **Comprehensive problem understanding:** Understanding the influence of bias in data science problems is fundamental to behavioral responsibility in data science. Data scientists should be aware of their own biases and how these biases affect the way they formulate the problem[173]. To counteract these biases, it is essential to involve diverse perspectives and stakeholders to develop a nuanced understanding of the problem and potential impacts on different groups. Beyond that, attention should be paid to examining historical data, and considering the historical context of the problem as it could reveal biases or systemic inequalities

that need to be addressed.

2. **Collecting unbiased data:** Imbalanced datasets can lead to biased models that perform poorly on minority classes. Data science practitioners may consider gathering more data for minority classes, oversampling minority classes, or re-weighting minority classes to address the issue [175]. Systems like Trifacta [1] enable dataset anomaly detection and quality assessment using quality rules such as data integrity constraints. Apart from that, consideration must be given to data points that do not yet exist in the data [77], which may result in a biased starting point.

3. **Careful data preparation:** Data preparation holds immense significance for behavioral responsibility in the field. This process involves cleaning and wrangling datasets to ensure that the data is accurate, complete, and free from bias [28, 173]. In addition to technically responsible techniques such as handling missing values, outlier detection and treatment, and thoughtful feature engineering, behaviorally responsible data preparation extends to the responsible handling of sensitive information, anonymizing data when necessary, and safeguarding privacy to uphold behaviorally responsible standards.

4. **Identifying biased interactions with data:** Identifying when bias may occur during analysis or interpretation, especially during interactive data analysis where the analyst may selectively look at certain data points while neglecting others (even though inadvertently) [100] is also a crucial step. Wall *et al.* [178, 179] propose an approach of computing and visualizing bias in user interactions during visual analysis.

5. **External Reviews for Accountability:** Accountability in data science should extend beyond technical reviews to include assessments by peers and stakeholders who will be impacted by the model. This involves treating the review process not just as a technical code review but also as a review of ethical practices and

implications. These reviewers can identify potential harm and unintended consequences that may not be evident to the technical team. One way to support this type of review is to support provenance tracking [31, 63, 192], so that data scientists may be held accountable not only to the outcomes of their models, but their process as well.

6. **Streamlining pipelines with checklists**: At the commercial level where stakes are typically high, checklists have been created to draw developers' attention to the entire pipeline specifically in machine learning-based data science [98, 44, 117]. Notable tasks within these checklists among many others, include ensuring fairness and privacy during data collection, transparency during analysis, and interpretability during inference.

Revisiting the smoking cessation example in Figure 3.2, a smoker may act responsibly by increasing his awareness of the health consequences of smoking to fight his addiction. Extending the idea to the machine learning example, a data scientist could act responsibly by evaluating the decisions of an ML model for potential downstream consequences (point 1 in the aforementioned list). In the visual data analysis example, responsible behavior could be to involve external evaluation of the visualizations to check if they convey information in a non-misleading way (point 5 above).

3.4 Operationalizing Behavior Change Theories for Responsible Data Science

In the preceding sections, we described the rich landscape of behavior change theories for data science. As we transition from a theoretical understanding to practical applications, it is essential to reflect on how these theories can be operationalized to design interventions in real-world data science scenarios. This critical step involves not only identifying and addressing specific behaviors within the context of data sci-

ence but also decomposing the design process of behavior change interventions in the context of data science environment. In this section, we aim to bridge this gap by offering a guide to translating theoretical insights into actionable steps.

In the previous section, behavior change theories were introduced chronologically based on the date of the publication of theories (e.g., factors of behavior change [72, 129, 127, 12], behavior change techniques [3, 130], and most recently work towards understanding mechanisms of action [34]). In this section, we alter this order to align with how we think about operationalizing these theories towards the development of interventions for responsible data science.

1. **Identify problematic and target behaviors:** It is crucial to pinpoint both problematic behaviors that might impede responsible data science practices and target behaviors that should be encouraged to replace them (see subsection 3.3.2 and subsection 3.3.3 for examples). This requires analysis of current methodologies and workflows. For instance, overlooking biases in data or algorithms can be considered a problematic behavior in the context of responsible data science, whereas actively seeking diverse data sources might be a target responsible behavior to encourage.

2. **Identify factors affecting problematic behaviors:** Building on the theories outlined in subsection 3.2.1, we need to identify various factors (*capability, opportunity, and motivation* [129]) that might influence the problematic and target behaviors identified in the previous step. We then need to assess whether digital interventions are appropriate given the factors involved. For instance, insufficient training can perpetuate undesirable practices in data cleaning, which might be rectified through interventions aimed at enhancing *capability*. Similarly, the lack of awareness among data science practitioners regarding the potential social impacts of their models can jeopardize the benefits to affected groups. This gap can be bridged by interventions that enhance their *motivation* to understand the

ethical consequences of the models they develop.

3. **Understand and employ appropriate Mechanisms of Action:** Once the factors affecting the problematic behavior are identified, the appropriate mechanism of inducing the target behavior needs to be identified and employed, as discussed in subsection 3.2.3. This involves understanding how different strategies leverage *capability, opportunity*, or *motivation* to initiate and sustain behavior change among data science professionals. Taking the machine learning scenario as an example (Figure 3.2), if data scientists lack *motivation* to commit more time to test model outcomes on different influenced groups, this could be bridged the mechanisms related to changing their *attitudes towards their behaviors* and updating their *beliefs about consequences* [34]. This not only helps designers to choose the most appropriate interventions for the digital context but also facilitates them to maximize impact.

4. **Envision potential interventions using BCT:** Having identified both the factors affecting problematic behaviors and the underlying mechanism of action, we can now envision potential interventions in the data science context by referring to the behavior change techniques [3, 130] introduced in subsection 3.2.2. These might include training programs, ethical guidelines, and decision-support tools that encourage reflection on the consequences of one's actions in the data science workflow. For example, to develop interventions in the data science environment that boost *motivation* by strengthening the understanding of ethical implications in the machine learning example (Figure 3.2), organizations can educate or train their data scientists, which facilitates their *beliefs about consequences* and *knowledge*. This *shaping of their knowledge* (as per BCTTv1[130]) could focus on identifying potential negative impacts of their models on target groups, using a variety of illustrative case studies.

Prior work has mapped factors affecting behavior change (FBCs) to specific interventions (BCTs) [164] and specific interventions to their underlying mechanisms of action (MoAs) [35]. To help designers choose appropriate FBCs, MoAs, and interventions, we provide a supplemental table that merges these mappings.

3.5 Interventions

In this section, we refer back to the two data science contexts in Figure 3.2 to apply these theories and discuss potential interventions for both the desired technically satisfactory and behaviorally responsible practices for the machine learning example in the second column (subsection 3.5.1), and visual data analysis example in the third column (subsection 3.5.2). Note that this is not an exhaustive account of interventions for these two contexts but merely describes some possibilities, grounded in behavior change theory. We further use this as an opportunity to describe these two examples as **usage scenarios** to explain how to apply the framework from Section 3.4, and in subsection 3.5.3, we provide our own **internal reflection** on the usage of this framework for envisioning behavior change interventions for responsible data science.

3.5.1 Interventions Designed for the Machine Learning Example

Maggie is a researcher who is designing Jupyter Notebook plugins to help people build more socially responsible models. She is collaborating with a data science team tasked with creating a loan approval model that avoids discriminating against potentially disadvantaged groups, such as female applicants [170]. Maggie decides to implement interventions based on our proposed framework.

She begins by **identifying problematic and target behaviors** within the team's workflow that could hinder the development of a fair model. Maggie is certain

that the team has sufficient expertise to tackle the technical challenges of data science models and deliver models with high accuracy. However she is concerned that they may not have enough understanding of how decisions made during the process of data wrangling and model building can have downstream effects, influencing the outcomes for different potentially disadvantaged groups. **Recognizing a lack of motivation as a significant factor**, Maggie wants to increase their awareness and empathy regarding the effects of their model decisions on socially disadvantaged groups.

To achieve this, Maggie thinks that she should **change the *attitudes towards their behaviors* (MoA) [34]** and employs this mechanism within her interventions. In the form of real-life stories of individuals, particularly female applicants who have faced repeated rejections from loan approval models, resulting in missed opportunities for housing or education, Maggie uses **sharing *social and environmental consequences* (BCT) [130]**. These stories are integrated into the team's data analysis environment within a notebook cell shown on the top, containing hyperlinks for these stories (Figure 3.3.a), serving as a constant reminder of the real-world impacts of their work.

Additionally, she **identifies *goals* (MoA) [34]** as another possible underlying mechanism and decides to employ ***goal priming* (BCT) [130]** intervention within the data analysis tools. She introduces prompts in the data analysis environment to ensure that data scientists are explicit about their goals throughout the workflow (Figure 3.3.b).

Enhancing the workflow in this way with contextual anecdotes and prompts to elicit development goals ensures that data scientists continuously reflect on the ethical aspects of their work. By envisioning potential interventions using our framework, Maggie ensures that her team is not only technically proficient but also behaviorally responsible.

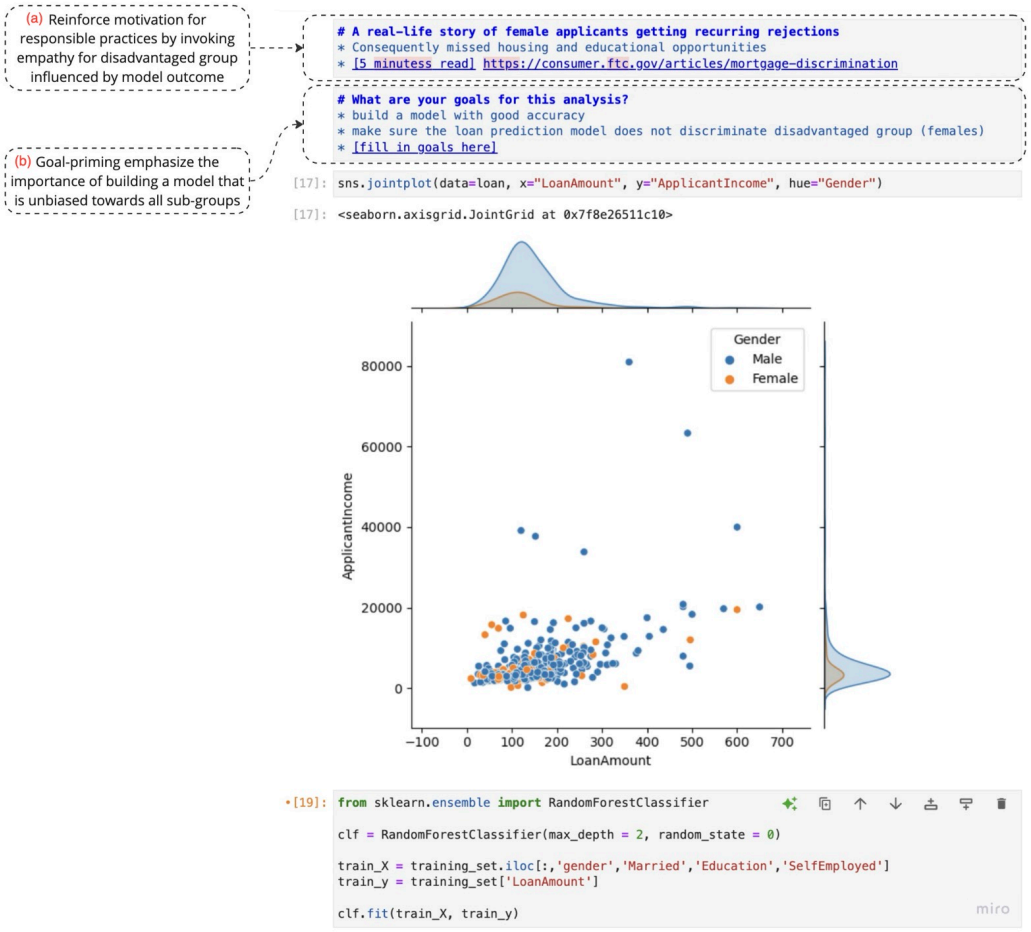

Figure 3.3: As data scientists start analyzing the loan approval dataset within a Jupyter notebook, this intervention (a) reinforces their motivation to practice responsible data science by sharing a real-life story that highlights the potential harm that model outcomes can inflict on disadvantaged groups, aiming to evoke their empathy; (b) follows-up with a goal-priming hint to emphasize the importance of behaving in an unbiased way towards vulnerable sub-groups that are influenced by the model's outcome.

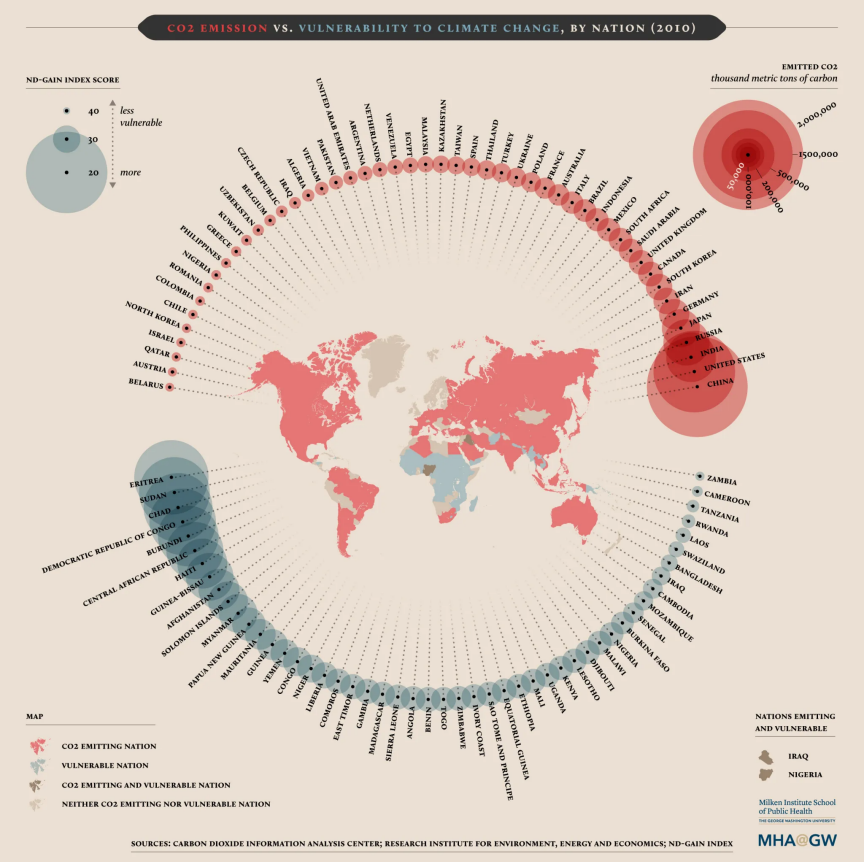

Figure 3.4: A data visualization showing which nations are major CO2 emitters, and which nations are vulnerable to the effects of these emissions. In its current state, this visualization might only help global policymakers like the Intergovernmental Panel on Climate Change (IPCC). By gathering feedback from viewer groups of different backgrounds like politicians, farmers, and students, this visualization could be made more effective by additionally visualizing how each group contributes to these emissions and how they could help alleviate the problem. Credits: https://onlinepublichealth.gwu.edu/resources/climate-change-emissions-data/

3.5.2 Interventions Designed for the Visual Data Analytics Example

Dylan is a quality control specialist within a data visualization team. His primary objective is to ensure that his team creates clear and trustworthy visualizations that are non-misleading and can be easily understood by people with various educational/occupational backgrounds.

For example, consider Figure 3.4 which estimates the carbon emissions from different countries, and highlights countries vulnerable to the effects of these emissions. The current version may only be helpful for policymakers to identify how they could address this problem, and may miss the chance to communicate with audiences such as students or farmers, who could address this problem in their own individual capacities. Realizing the scope to maximise impact, Dylan decides to design an intervention tool based on our proposed framework to help the visualization team.

While **identifying problematic and target behaviors** in the visualization workflow, Dylan realizes that although technically adept, the designers in his team might not realize how many different target audiences may encounter their visualizations. He considers this problem in terms of a textbflack of motivation, and decides to use the *social influences* MoA [130] to help them realise the potential impact on society. Dylan designs interventions to prompt designers to estimate the anticipated range of their target audience [111] (policymakers, urban public, university students, rural public,). He generates possible scenarios to expose designers to diverse questions audience members might ask about the visualization; how different agricultural activities contribute to these emissions might interest rural communities; students might be interested in the steps they can take to help alleviate this problem or increase awareness.

Alternatively, Dylan also **recognizes a lack of opportunity as a factor**. He understands that although the designers are aware of the different target audience groups, not getting feedback from these diverse groups is hindering them from creating more inclusive and effective visualizations.

To achieve this, Dylan **incorporates *feedback processes (MoA)* [130]**. Using our framework, Dylan figures out that gathering feedback from different target audience groups, which falls under ***feedback of behavior* (BCT) [130]** could be employed. Dylan thus generates customizable annotations that appear upon completion

of visualizations along with a shareable link. These annotations nudge the designers to proactively communicate and gather feedback from stakeholders from diverse backgrounds, capturing their individual perspectives and levels of comprehension regarding the visualizations, and facilitating the designers in pinpointing potentially confusing or divisive areas in need of improvement.

3.5.3 Internal Reflection

In this section, we reflect on the usage of our proposed framework for envisioning behavior change interventions for responsible data science. We do so by reflecting on some explicit questions.

Where was the framework the most helpful? As seen in the previous two subsections, the framework helped Maggie and Dylan enlist multiple possible FBCs, and the possible MoAs and corresponding BCTs to bring about the desired behavior change. Maggie used the framework to identify different MoAs (*attitude towards behavior* and *goals*) to better motivate the data scientists to consider the downstream effects of their models. On the other hand, Dylan found two different FBCs - *motivation* and *opportunity* and accordingly employed BCTs to help visualization designers. The framework thus acted as a comprehensive, though not necessarily exhaustive tool, to generate ideas in a systematic way, without which both Maggie and Dylan could have missed out on potential additional ways to bring about responsible behavior change.

Where was the framework the least helpful? The framework provides a consolidated space of possible approaches for identifying the scope for responsible behavior change, and actionable techniques to bring about the change. However occasionally, the boundaries between the individual FBCs, MoAs, and BCTs that are applicable in a situation are not very clear. For example, *attitude towards behavior* MoA could be employed through BCTs of both *consequences* and *rewards*. The

appropriate alternative is evident based on the context in most cases, Maggie used the MoA to inform data scientists about the *consequences*. However, the blurred boundaries or redundancies between some of these terms might cause difficulties for practitioners to use the framework.

Further, there is an open-ended nature in the interpretation of a certain situation as lacking a certain FBC. For example, Maggie's intervention of providing *prompts for goal priming* could boost *motivation* of the data scientist to also address downstream consequences of their models. However, it could also be interpreted as providing *opportunity* to the data scientist through *prompts/cues* during model development to inform them about downstream consequences. The latter interpretation assumes that the data scientist is already motivated but lacks the right opportunity to be reminded about responsible behavior.

Although these limitations of our framework might create complications in choosing the appropriate BCT, the framework also helps practitioners by making them aware of the possible multiple interpretations of the situation. We thus see this framework as providing a full range of behavior change solutions, while leaving the responsibility to choose the right alternative to the practitioner.

3.6 Discussion

In this project, we introduced a new perspective on responsible data science that elevates the importance of responsible *agents* through the lens of behavior change. Our research complements existing work in guideline and curriculum development by encouraging analysts to adopt more responsible analysis behaviors through their real-time interactions with data science tools. Based on this novel perspective, we identify pressing research challenges moving forward.

3.6.1 Challenge 1: Intervening at the Right Time

When introducing interventions to foster responsible data science practices, it is crucial to strike a balance where interventions are neither absent when assistance is needed nor persistent to the point of causing frustration. Hence, the timing and appropriateness of interventions are crucial not only for their effectiveness but also for ensuring a positive user experience. Adopting the concept of *triggers* for behavior change interventions [72], we could conceptualize *when* to initiate an intervention as a *disruptor*. We present several heuristic approaches to *disrupt* data science practices to initiate an intervention, hoping to inspire potential solutions to this open challenge:

Disrupt by Data Science Phase. The output at each phase of a data science workflow serves as input to, or otherwise influences, the next phase. A seemingly benign negligence of technically satisfactory or behaviorally responsible practices in one phase can significantly impact downstream deliverables through the propagation of inaccuracies and biases. Thus, the beginning or end of a phase in the data science pipeline may be an apt time to disrupt the user's process to intervene.

Disrupt by Algorithmic Performance. Interventions could also be designed to disrupt a data scientist's practices based on active monitoring of metrics. In a fairness-aware data science project, crime recidivism analysis, algorithmic bias is treated as an important evaluation metric. One disruptor, in this case, could be to continuously monitor fairness metrics that have been identified as critical for the task to identify important dips to the metrics of interest when data is transformed, model parameters are tuned, etc.

Disrupt by Minimal Timeline. Another possible disruptor could be related to minimum expectations for time spent on some tasks. The underlying assumptions are that (1) some tasks must be completed, running fairness checks on data prior to model building, and (2) there is an expected amount of time to complete various tasks, which when below a threshold, may be indicative of negligence to fully understand the data

or model. For instance, if an analyst does not dedicate time to exploring the data prior to creating a model from it, it could lead to unknown problems downstream.

Disrupt by Third-party Review. Interventions could also involve an external ethics review board or third-party auditors to conduct periodic assessments of the data science project to provide an unbiased evaluation of responsible data science practices. Furthermore, interventions can be based on feedback from stakeholders, end-users, or affected communities to address any ethical issues or concerns that arise during the project's lifecycle.

Disrupt by Programming Specification. Disruptors can also be identified through established issues in technical standards. For example, a normative coding style is beneficial to avoid ambiguity for implementer reading, collaboration purposes, and future model maintenance. Non-normative coding and variable naming habits may impede collaborators from easily verifying the code which can discourage or hinder future checks on responsible practices. Thus, disruptors could identify violations of established programming patterns.

3.6.2 Challenge 2: Facilitating Lasting Behavior Change Through In-The-Moment Interventions

In this project, we focus primarily on settings and examples for in-the-moment interventions that potentially result in short-term positive behavior changes during data analysis. However, it is unclear *if* and *how* these interventions will fundamentally change the long-term practices of analysts [143]. We view in-the-moment interventions as a subset of behavior change techniques (BCT) for facilitating rigorous data science. The superset also includes interventions for long-term behavior change or habit formation, provide learning materials like enrolling in a course to learn new skills, checklist generation (Planning and Repetition), setting goals and tracking progress [109]. Alternative theories and applications of behavior change interventions

emphasize settings intended to encourage longer-term habit formation [68, 163]. Accordingly, we observe at least three major challenges in establishing lasting behavior changes in analysts.

Ensuring smooth hand-offs between in-the-moment interventions for short-term behavior change, and interventions for long-term behavior change will hopefully expose analysts to a range of experiences to reinforce rigorous data science practices. Examples include pointing the analyst to relevant online tutorials or courses on statistical testing (long-term) *after* detecting improper statistical testing being performed and recommending more appropriate tests (short-term).

Accounting for the evolution (or devolution) of the analyst. The analyst's practice of data science may change over time, which may influence both the efficacy of interventions and disruptors. For example, disrupting the flow of a confident analyst during a well-defined task may hinder rather than accelerate their work [93]. How then do we interrupt an analyst who is initially receptive to interventions, but starts ignoring them later for an unknown reason?

Tackling Long-term Bias. Another critical aspect of designing long-term interventions is considering the potential long-standing biases that may exist or develop over time in data scientists and analysts, independent of their use of interventions. These biases could influence their decision-making processes and perpetuate existing inequities[78, 124], which may be less responsive to intervention. Thus recognizing the boundaries of where interventions may be effective is important for designing more ambitious interventions.

3.6.3 Challenge 3: Measuring Efficacy & Boosting Adoption

We hypothesize that a collection of complementary evaluation techniques will be needed to understand the complex interplay between system behavior and user behavior when measuring behavior change in data science.

How do we measure the efficacy of deployed interventions? Are the same metrics used to choose a disruptor and an intervention sufficient to understand their efficacy? It also becomes crucial to isolate whether the cause of positive behavior change is indeed the intended intervention, or attributable to some other confounding factor. Furthermore, would repeated measures of the same heuristic over time provide sufficient information to show progress? Or do we need to measure specific long-term outcomes[78]? Incorporating these long-term fairness considerations can provide a more comprehensive view of the effectiveness of behavior change interventions. For instance, tracking the long-term outcomes of interventions on loan approval fairness can reveal whether initial improvements in fairness metrics translate into sustained equitable lending practices.

3.6.4 Challenge 4: Incentives Versus Consequences to Induce Behavior Change

Encouraging positive behaviors or punishing negative behaviors is analogous to a carrot versus stick metaphor. The examples we emphasize in this project primarily focus on positive reinforcement (carrots). However, these so-called "carrots" are not the only way to encourage responsible data science practices. Alternatively, how to establish consequences (sticks) operationalized into interventions as a way to *enforce* course correction has yet to be explored. The BCT taxonomy [130] identifies relevant interventions in the categories of Reward and Threat and Scheduled Consequences which target the capability of the analyst through behavioral regulation or by changing their attitudes towards the behavior. However, it is unclear what role data science tools should play in holding analysts accountable for their contributions to irresponsible data science outcomes. For example, how do we infer the scope of an analyst's contribution to a certain outcome, positive and/or negative? Once this scope is established, how do we reason about the consequences of an analyst's contributions in

relation to the final outcomes?

3.6.5 Challenge 5: Automated Versus Behaviorally Responsible Data Science

There exists a tension between automation and behavioral responsibility. For example, autoML techniques aim to reduce the reliance on analysts for making design decisions towards creating satisfactory models [104]. While these methods reduce the analyst's time and effort in generating satisfactory models, autoML methods are poorly designed to support human oversight and agency within this process [45]. With a reduced ability to intervene in the model design process, the analyst's behavioral responsibilities may clash with the goals of autoML systems. Further investigation is needed to understand how behavioral responsibility can meaningfully engage with highly automated data science tools.

3.6.6 Challenge 6: Enhancing Education and Training for Data Science Practitioners

Throughout this project, we highlight the importance of education in promoting responsible data science practices. Echoed by many prior works in this space [13, 161, 29], one potential direction for the responsible data science research community is to delve deeper into developing comprehensive educational frameworks and training programs that equip the current and future data science practitioners with the necessary skills and ethical mindset to navigate complex data science environments. These programs should go beyond technical proficiency to include modules on ethical reasoning, bias detection and mitigation, and the societal impacts of data science decisions. Additionally, integrating behavior change theories into training curricula can help instill long-lasting responsible behaviors. Research should also explore in-

novative teaching methods, such as experiential learning[177], case studies[116], and interactive simulations[157], to enhance the learning experience. By advancing education and training, we can prepare data scientists to not only excel technically but also to act responsibly and ethically in their professional roles.

3.7 Limitations

While this chapter establishes a theoretical foundation for applying behavior change models to responsible data science, several limitations must be acknowledged. First, the translation of psychological frameworks from health and environmental domains to data science involves inherent differences between the domains which make it non-obvious how appropriately these frameworks can suit this new context. Data science workflows differ significantly from personal health behaviors in their complexity, collaborative nature, and professional context. The behavioral models we have adapted may require further refinement to fully capture the nuances of data science practice.

Second, our theoretical approach intentionally separates technical and human factors for analytical clarity, which risks oversimplifying the deeply sociotechnical nature of data science work. In practice, responsible data science emerges from the complex interplay between human decision-making, organizational structures, and technical systems. Future work should reintegrate these dimensions through empirical studies of how data scientists navigate ethical considerations within existing technical constraints.

Third, the efficacy of behavior change interventions in data science remains largely theoretical at this stage. While we have drawn parallels to successful interventions in other domains, the data science context presents unique challenges that may limit transferability. As subsequent chapters of this dissertation will demonstrate, de-

signing and empirically validating interventions requires moving beyond theory to implementation and evaluation in realistic data science scenarios.

3.8 Summary

In this project, we addressed **RQ 1** (*How can I utilize behavior change theories to inform a novel framework for responsible data science using the lens of behavior change interventions?*) by introducing the concept of behavior change interventions for data science, where we focus on data science behaviors as possible predictors of biased outcomes. We first synthesized a definition of responsible behaviors in data science work for both humans (behavioral) and systems (technical). Secondly, we illustrated how existing psychological models can inform the design of behavior change interventions in data science contexts. Lastly, to inspire the design of potential interventions, we presented concrete examples of possible interventions aimed at encouraging socially responsible behaviors within the data science context. We concluded this project by describing the next action item uncovered by this vision project and calling on our community to explore this new research area of behavior change interventions for responsible data science.

Chapter 4

Synthesizing a Design Space of Behavior Change Interventions for Responsible Data Science

4.1 Motivation

Behavior profoundly impacts the development and deployment of data science models. In Chapter 3, we established a conceptual framework of responsible data science using behavior change theories to address **RQ1** and proposed high-level four-step guidelines for designing behavior change interventions (Chapter 3.4). While this foundation provides a valuable starting point, translating these principles into effective interventions requires a more comprehensive and systematic design framework.

This chapter introduces a 5W1H (Who, What, When, Where, Why, and How) interrogative design space that bridges theoretical understanding and practical application for responsible data science interventions. This structured approach is crucial because designing effective interventions requires navigating complex interdependencies between technical systems and human behaviors. Without systematic guidance,

intervention designers may overlook critical factors or fail to consider the full range of available strategies, potentially leading to ineffective or counterproductive outcomes.

A well-defined design space creates a common language and conceptual structure that enables knowledge accumulation across different contexts, accelerating collective progress in responsible data science. After careful consideration of alternative approaches, we deliberately chose the 5W1H interrogative framework for several key advantages: it is domain-agnostic yet easily adaptable to responsible data science contexts; it ensures coverage of all essential design dimensions; and it employs familiar questioning patterns that practitioners already use in their work. The framework's natural interrogative structure aligns well with design thinking methodologies widely used in human-computer interaction [91, 160, 165, 166], making it accessible to interdisciplinary teams. Additionally, this approach offers flexibility for future expansion and refinement as the field of responsible data science continues to evolve.

This project addresses the second research question: *How can we scaffold the design and development of behavior change interventions for responsible data science? (Goal II: Synthesize Design Space).* The objective is to outline principles that can guide the design of interventions in the context of responsible data science. To guide our design space, we adhere to the 5W1H interrogative framework[85], encompassing **Who**, **What**, **When**, **Where**, **Why**, and **How**. We demonstrate the overview of the proposed design space in Figure 4.1.

1. **Why** do you as a designer want to intervene?
2. **Who** is the target of the behavior change intervention?
3. **What** key objectives does the intervention seek to influence?
4. **When** is a suitable time to intervene?
5. **Where** do these interventions take place?
6. **How** can we design effective interventions?

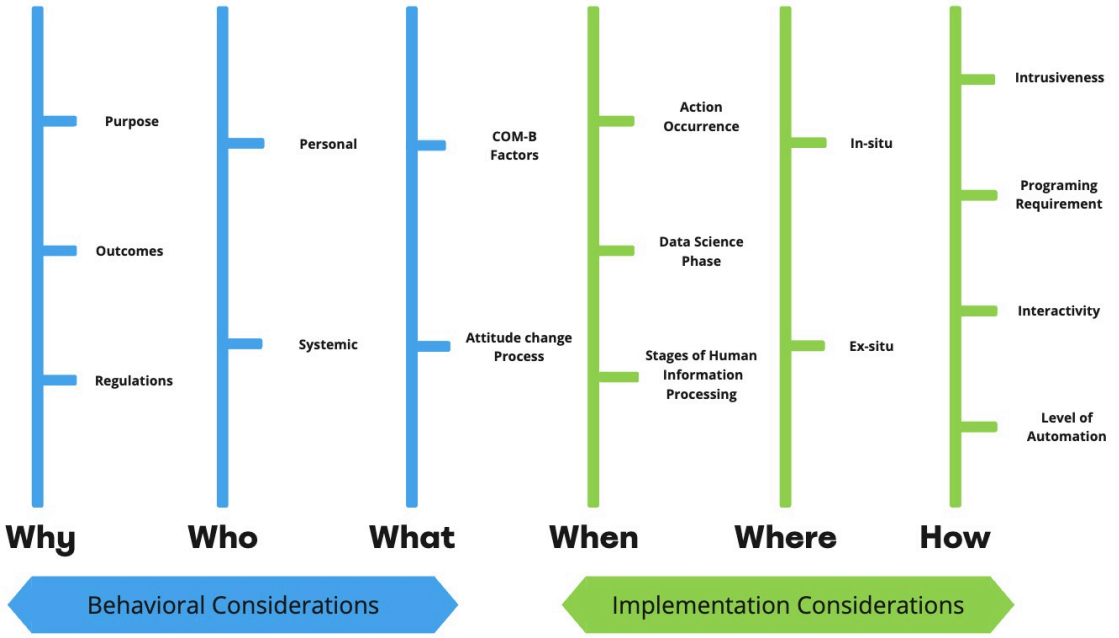

Figure 4.1: An overview of 5W1H design space proposed in the work.

Furthermore, we survey and characterize existing responsible data science tools to validate the coverage of this design space. Through this analysis, we validate the relevance and applicability of the design space, and identify future research opportunities where current tooling falls short. In summary, this section makes the following contributions:

- We introduce a **design space of behavior change interventions** to promote responsible data science practices, comprised of *behavioral considerations* and *implementation considerations*.
- We present a complementary **interactive website** for convenient use of the design space by potential intervention designers.
- We validate the breadth and applicability of the design space through a **qualitative analysis of 23 data science tools** and demonstrate its potential with two **usage scenarios**.

This design space fills two key gaps in the literature. First, existing frameworks

for responsible data science often focus on *practitioners* by providing checklists or guidelines to follow [150, 149, 134, 75]; yet, these resources lack actionable strategies for *tool developers* who aim to promote behavior change through the development of interventions. Our framework is thus complementary to these efforts, offering a flexible, structured, and actionable approach to fostering ethical responsibility in tool design and development. Second, this design space enables us to move beyond static references for compliance, and instead supports researchers and developers to *translate good practices into actionable applications*. This project has been accepted by ACM IUI 2025 as a full paper[53].

4.2 Design Space Rationale

The decision to utilize the 5W1H framework [85] in designing behavior change interventions for responsible data science stems from its versatility and widespread applicability across various domains. The 5W1H approach—encompassing Why, Who, What, When, Where, and How—provides a comprehensive, yet structured way to navigate the complexities of behavior change interventions by addressing key questions that guide the design process. These dimensions are grounded in a robust basis in the sciences [167, 79, 89] as well as recent applications in Human-Computer Interaction (HCI) and Visualization [91, 160, 165, 166]. To ensure clarity, we have divided these dimensions into two categories:

- **Behavioral considerations** (Why, Who, and What) which focus on understanding the motivations, audience, and targeted behaviors
- **Implementation considerations** (When, Where, and How) which deal with the practical aspects of timing, context, and delivery methods

It is important to note that these dimensions are not strictly orthogonal but represent complementary perspectives that work together to drive responsible behavior

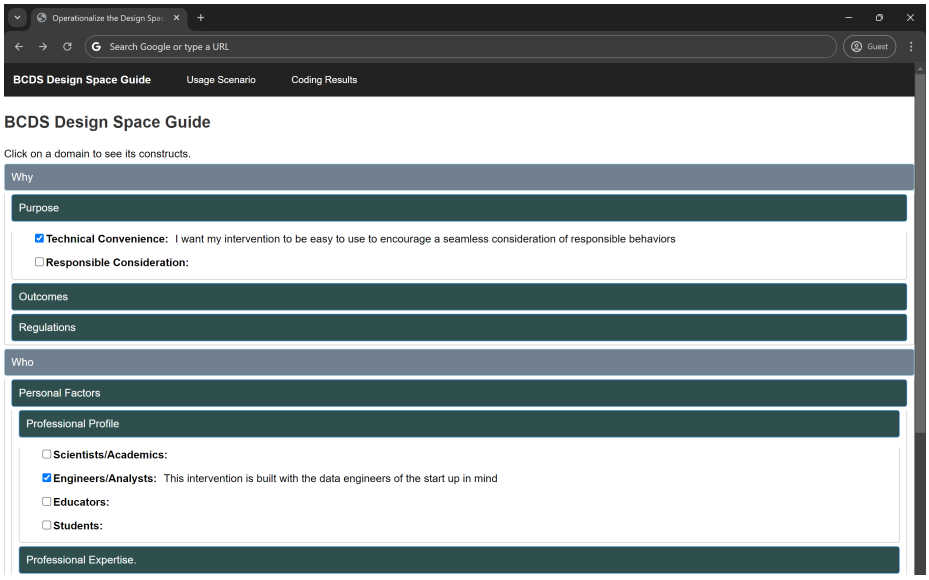

Figure 4.2: A screenshot of the interactive BCDS design space website.

in data science.

Interactive Website. To facilitate the exploration and application of the design space, we developed an interactive website. The website allows users to step through dropdown sections representing each branch in the design space. Users can decide which aspects are needed for their intended intervention and enter notes or annotations throughout each subsection (see Figure 4.2). Once the user has explored the design space, the completed design and associated notes are generated into a downloadable PDF. The designer can refer to the document and share it with collaborators throughout the design process. The full interactive website source code and completed scenario examples can be found within the supplementary materials.

Usage Scenarios. We additionally contribute two usage scenarios. The usage scenarios are meant to illustrate how our design space supports an intervention designer in achieving a focused vision for their tool. We show two illustrative examples in Sections 4.3.4 and 4.4.4. The first demonstrates how the design space can be used in the early stages to systematically understand the user context and subsequent goals for the intervention. The latter is aimed at the ideation phase, where the user context

is known and interventions are ready to be designed.

The design space is intended to be a living, collaborative and instructive artifact. Therefore, it can be updated, shared, and referenced for a variety of activities such as documentation, informing stakeholders (especially non-technical), processing design feedback and aiding software developers in creating the intended tool. In addition to the two usage scenarios we outline, we envision the design space can be used in other contexts as well, e.g.:

- to develop internal tools and guidelines to help company ethics advisors figure out the best way to ethically guide company data scientists
- to design an interview study to better understand the needs of a target data science community that you are building an intervention for
- as an evaluation tool for intervention designers to determine if additional features should be added or existing features should be refined

4.3 Behavioral Considerations

In designing responsible data science interventions, it is essential to understand the human and behavioral factors that influence ethical practices. This section introduces the behavioral aspects of intervention design by exploring:

- *Why do you as a designer want to intervene*,
- *Who is the target of the behavior change intervention*, and
- *What key objectives does the intervention seek to influence.*

Furthermore, we demonstrate in subsection 4.3.4 how this design space can be applied through a usage scenario that illustrates how the behavioral dimensions can facilitate the development of a responsible data science intervention tool.

4.3.1 Why: Why do you as a designer want to intervene?

Understanding the motivations behind behavior change interventions is fundamental to designing strategies that are both effective and sustainable [37]. The "Why" dimension explores the driving forces that compel the adoption of ethical practices in data science. We characterize three broad categories that represent different perspectives on motivations for a developer to intervene in data scientists' practices: **Purposes**, **Outcomes** and **Regulations**.

Purposes

Purposes characterize our reasons behind intervening, which we further categorize into two areas: technical convenience and responsible considerations, as described next.

1. **Technical Convenience**: We could intervene to simplify the process of implementing responsible data science practices through streamlining or automation. For example, an intervention might automatically detect missing data and recommend pre-processing techniques to assess the fairness of different strategies for dealing with the missing data (e.g., imputing missing values vs. discarding the data [52]). Similarly, a plugin could suggest encryption methods during data export to make compliance with data security standards easier for the user without requiring extensive manual configuration.

2. **Responsible Consideration**: We could also intervene to incentivize data scientists to engage in ethical practices in general. This could be achieved by highlighting the long-term benefits of ethical actions, such as improved model accuracy and public trust. For instance, an intervention might prompt users to assess the social consequences of their model by providing a pop-up message highlighting potential bias and fairness issues affecting underrepresented groups.

Outcomes

Thinking about the "Why" dimension in terms of outcomes is crucial for measuring intervention success and ensuring that it leads to a targeted or meaningful behavior change. We can think of this dimension according to *promoting* target outcomes or *hindering* problematic outcomes:

1. **Positive Behavior Promotes Outcome**: Interventions designed to promote positive behaviors encourage actions that lead to beneficial outcomes in data science projects. For instance, one possible intervention could be a real-time bias monitoring tool that reminds users to refine the model configurations when their model's outputs show potential bias against certain groups. This tool could guide the user through steps to adjust the model or provide resources on alternative algorithms or techniques. The direct outcome is a more equitable treatment of individuals by the models developed, which upholds ethical standards and improves societal impact.
2. **Negative Behavior Hinders Outcome**: Interventions aimed at reducing negative behaviors focus on preventing actions that could lead to harmful outcomes. For example, by integrating features that track and report the use of data in unauthorized ways, an intervention could alert administrators or data ethics officers if sensitive data is being misused. Another example could be mandatory review checkpoints before a model is deployed, preventing models from not being evaluated for ethical compliance and bias.

Regulations

Regulations focus on the need to align data science practices with both external regulations and internal standards, ensuring that individual actions are bound by ethical, legal, technical and organizational mandates. Regulations can be further

categorized into at least three relevant types:

1. **Technical Standards**: These are specific, often quantifiable standards that models and data handling procedures must meet, such as reaching certain confusion or fairness metrics. Interventions here might involve compliance checks integrated into data science platforms that automatically verify whether the data management, model implementation and development processes meet established technical benchmarks for security and efficiency.
2. **Legal Standards**: Legal requirements demand adherence to laws and regulations, such as GDPR [176] for data privacy in the European Union or HIPAA [4] in the United States for health data. Interventions could include compliance modules within tools like Jupyter Notebook that guide data scientists through necessary legal documentation and ensure that their work complies with relevant laws.
3. **Ethical Standards**: These standards reflect the moral obligations of the profession and are often guided by broader ethical principles of harm prevention and fairness. Interventions could consist of ethical audit trails in software that document decision-making processes and flag potential ethical issues, prompting users to reconsider decisions that may have harmful implications.

4.3.2 Who: Who is the target of the behavior change intervention?

Interventions that influence behavior must be personalized to their audience to be effective [158]. The "Who" dimension addresses the diverse spectrum of individuals and groups involved in data science processes. This differentiation is crucial because data science is not a monolithic field [173]; it involves various stakeholders with different roles, expertise, and influence over data-driven outcomes. Categorizing

the target audience in different dimensions can ensure that interventions are not only appropriately designed but also contextually relevant in order to increase the likelihood of adoption and impact [156]. This section characterizes the target user along two complementary dimensions: **Personal Factors** and **Systemic Factors**. These factors may help inform optimal behavior change interventions that are effective for specific types of users.

Personal Factors

The Personal Factors dimension includes factors related to the individual characteristics of data science practitioners. We describe three potentially useful ways of thinking about characteristics of target users:

1. **Professional Role:** Different professional roles entail varied responsibilities and influence within data science projects, which suggests a need for customized interventions designed for different professional profiles. For instance, **Scientists/Academics**, **Engineers/Analysts**, **Educators**, and **Students** all have very different relationships with data science practices. For example, engineers need real-time tools that can detect and mitigate biases in their model implementation, while educators could benefit from interventions that facilitate their data science teaching process or interactive tutorials that provide engaging learning experiences to their students.
2. **Professional Expertise:** Expertise level influences how interventions are received. Data science projects not only include knowledge of data science broadly, but also require fundamental knowledge of the target task domain. Hence, two specific areas of expertise that are especially important for making informed choices of behavior change interventions include **Data Science Expertise** and **Domain Knowledge**. Those lacking in data science expertise could benefit

from interactive tutorials that introduce core data science concepts along with ethical considerations. On the other hand, data science professionals who lack domain-specific knowledge could benefit from interventions that provide domain-specific guidelines and best practices for ethical data handling and analysis.

3. **Personal Profile:** This dimension emphasizes the individual characteristics of data scientists who will use the interventions, focusing on their own identity and role within the data science process. Below we exemplify some significant aspects of the intervention users that the intervention designers should take into consideration, including **Gender**, **Ethnicity**, **Age Groups**, and **Personality Traits**. For example, individuals' gender identity and unique experiences can influence how they interact with technology and perceive ethical issues. An intervention could include gender-sensitive training modules that highlight common biases in data science practices and offer strategies to overcome them. Furthermore, different age groups may have varying levels of familiarity with technology and ethical norms. Younger data scientists might be more comfortable with interactive, tech-driven interventions, while older professionals might prefer traditional methods. Interventions should cater to these preferences.

Systemic Factors

Systemic Factors reflect the broader context in which professionals operate, including the organizational and cultural norms that influence their work [8]. Unlike individual behaviors or personal intentions, systemic factors acknowledge that disparities and biases can emerge unintentionally due to the workings of larger social, organizational, or technological systems. Interventions designed with an understanding of these systemic factors can better align with existing workflows and cultural norms, thereby enhancing adoption and effectiveness. We describe two potentially relevant perspectives on systemic factors:

1. **People**: This category acknowledges the diverse range of stakeholders involved in or affected by data science projects. Relevant factors include **Data Privilege** and **Collaborative Factors**. Data privilege indicates the accessibility to data based on one's position or role. An intervention could highlight these disparities by alerting users when their dataset includes proprietary information unavailable to others, encouraging them to consider whether this advantage might unintentionally contribute to bias or inequity in their model's outcomes. Collaborative factors focus on how collaborative dynamics influence data practices (e.g. Multi-discipline collaboration and team culture). Interventions might feature collaborative coding tools or shared Jupyter Notebook environments that encourage transparency, peer review, and the ethical sharing of insights and methodologies.
2. **Organizational Process**: Organizational processes govern how data science work is conducted. This category is split into **Process Orientation** and **Project Clarity**. Process orientation refers to the overall approach an organization takes toward data science projects including the specific workflows, priorities, and methodologies it adopts. Interventions could include automated workflow tools in Jupyter Notebook that ensure ethical checkpoints or reviews are a routine part of all data science projects. Project clarity ensures that all team members have a clear understanding of project goals and ethical guidelines. A Jupyter Notebook extension could, for example, provide project dashboard functionalities that explicate project roles, expectations, and ethical considerations at each stage of a project.

4.3.3 What: What key objectives does the intervention seek to influence?

The "What" dimension focuses on identifying the behaviors that the intervention aims to modify or reinforce. Understanding which behaviors to target is crucial

for designing interventions that can effectively guide data scientists toward more responsible practices. Additionally, we consider attitude change in this section because shifting attitudes can lead to more sustainable and internalized behavior change [120]. To consider what behaviors we aim to change, we orient this dimension with two fundamental questions: **What behavioral factors (COM-B) [129] are being addressed?** and **What attitude change processes [105] are being addressed?**

What Behavioral Factors Are Being Addressed?

The COM-B model [129] offers a framework for understanding Behavior (B) as a function of three core factors: Capability (C), Opportunity (O), and Motivation (M). The factors that influence a user's behavior the most may vary across different scenarios. Identifying which of these factors may need to be bolstered could help make the intervention more effective:

1. **Capability**: This refers to an individual's psychological and physical capacity to engage in the behavior. If capability is lacking, interventions can focus on ways to enhance it accordingly. For example, an intervention could be an embedded tutorial or pop-up hints that guide users on how to implement data privacy measures or check for data bias.
2. **Opportunity**: This involves all the factors that make the behavior possible or prompt it. Effective interventions should help create opportunities for responsible data science practices to take place. For example, an intervention might modify the Jupyter Notebook interface to make ethical guidelines more accessible or to facilitate discussion and peer review before publishing results.
3. **Motivation**: This refers to the brain's processes that energize and direct behavior, which can be reflective (planning, evaluating) or automatic (habits, emotions) [73]. An intervention might include motivational reminders or gamified

elements that reward users for consistent application of ethical practices, thereby boosting motivation.

What Attitude Change Processes Are Being Addressed?

In addition to behavioral factors, behavior change interventions in responsible data science also need to address attitude change processes. We draw on three well-established attitude change processes from Kelman [105]: compliance, identification, and internalization to contextualize them in responsible data science. Short-term behavior change, such as compliance, tends to be externally motivated, often driven by rewards or penalties—a metaphorical "carrot and stick" approach. On the other end of the spectrum, long-term behavior change involves internalization, where the behavior becomes inherently motivated and aligned with personal values, leading to more sustainable ethical practices.

1. **Compliance:** This refers to the influence that is accepted in order to avoid punishments or gain rewards, often occurring when behavior is monitored or under surveillance [105]. Compliance typically drives short-term behavior change, as data scientists may comply with data privacy regulations, such as GDPR [146], to avoid legal penalties or reputational damage.
2. **Identification:** This occurs when individuals adopt behaviors or attitudes because they aspire to emulate someone they admire or respect [105]. In responsible data science, identification can be leveraged by promoting role models within the field who exemplify ethical behavior. For instance, sharing highlight stories from senior data scientists, professors, or prominent figures in the field who advocate for fairness, transparency, and ethical practices can inspire others to follow their example.
3. **Internalization:** Internalization is the deepest form of attitude change, where individuals adopt behaviors because they align with their personal values [105].

This process is associated with long-term behavior change, as data scientists follow ethical guidelines out of an inherent belief in the importance of responsibility. Interventions aimed at fostering internalization might focus on education and awareness-raising efforts that connect ethical practices with personal values. For instance, providing informational links that explore the societal impacts of biased models or the long-term consequences of data privacy breaches can help data scientists understand the moral imperatives of their work.

4.3.4 Usage Scenario: A State Government's COVID-19 Support Model

Intervention Inception

The Georgia Department of Economic Development was awarded a federal grant to support small businesses adversely affected by the COVID-19 pandemic. More specifically, the grant is focused on assisting businesses owned by minorities, women, veterans, immigrants, first-generation immigrants, individuals with disabilities, or identified members of the LGBT+ community (classified as "protected groups"). The department has access to the data of state registered small business within the past four years. The team assembled to implement the grant project decided to build a model to determine if a business is eligible for the funding and how much they should receive from the available funding. A small, contracted data science team is brought on to develop the model. The fiscal and political experts from the original team are responsible for providing advice and evaluating the model. The lead of the technical team, Sean, wants to create a responsible data science intervention to ensure the funding algorithm equitably allocates funding opportunities across all of the groups of interest. The technical lead decides to use the Behavior Change for Responsible Data Science design space to determine the direction of the intervention.

Key Insights from Design Space

Sean uses the interactive Behavior Change in Data Science website to annotate notes about the dimensions he finds helpful. Sean found the "Systemic Profile" of the **"Who"** branch to be an instructive way to clarify the organization of the team. There is a healthy multidiscipline collaboration between the subject matter experts in the department and the data science consultants. Sean values the input of the financial and political experts and knows that the model has to be signed off by the experts before it is deployed. The data scientists report to Sean, and Sean works with the department experts to get feedback and transform the feedback into technical tasks. Scrolling down the webpage, the "COM-B factors" in the **"What"** section helped Sean clarify the main goal of the intervention: to improve the opportunities for the data science team to review how closely their work aligns with primary responsibility goals. The **"Why"** branch spurred Sean to seek answers from the legal expert of the team. He understands that there is a strong ethical push to the project, but he is unsure of the legal standards and regulations that the data scientists should be aware of. After reviewing the **"Why"** section, Sean communicates his queries to the legal expert, who hosts a meeting with the technical team to outline all the relevant regulations and government laws the team needs to consider for the project. All in all, Sean's exploration of the behavioral considerations of the design space helped him identify the social dynamics he wants the intervention to support and the gaps of knowledge he needs to address before moving forward with the intervention and project in general. After completing the behavioral considerations, Sean reviewed the implementation considerations of the design space to complete the design space and hit the "Download" button to save the annotations for future design usage.

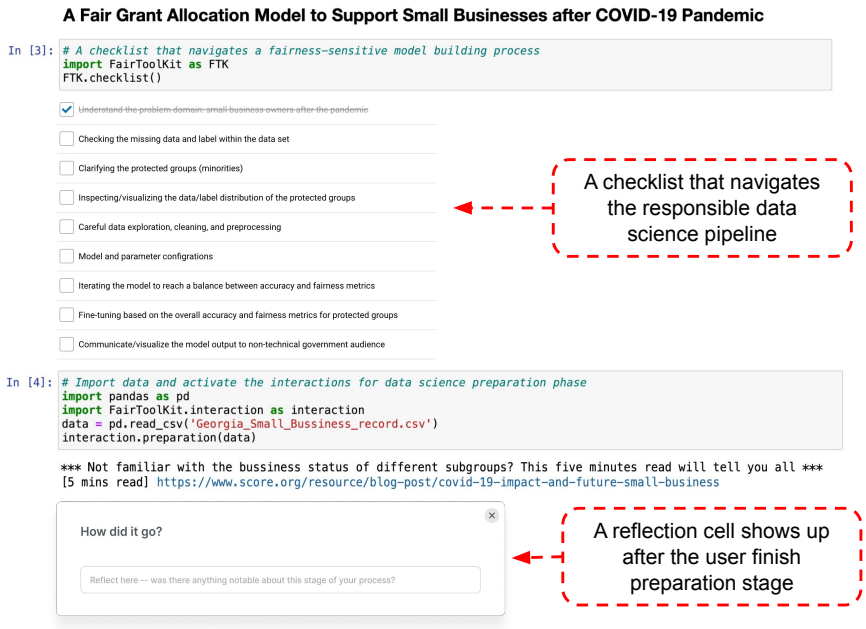

Figure 4.3: An exemplary intervention Sean envisioned.

Design Space Impact

Sean downloaded his completed report from the website and added it to the project folder. In the first team meeting with the department and technical team, he printed copies of the report and shared it with the team as a part of the meeting material. The department team appreciated the detailed focus on equity and the technical team appreciated the guidance the tool would provide. With the report as a guide, Sean led the technical team through the first sprint to create a simplified version of the intervention. This version provides a static checklist and an interactive cell that enable data scientists to reflect on their process and potentially recognize flaws for each stage in the data science process (Figure 4.3). The technical team found the design of the intervention very helpful to clarify the ethical goals of the project at each stage of development. Sean presented the intervention tool to the subject-matter experts in the subsequent meeting to get feedback on the accuracy of the checklist content. The department team was very excited by the reflection feature because it could be used as qualitative data to record progress to their grant funders. They also

asked for a digital copy of the report to add to their documentation as well. Using the intervention, when Sean met with the department team each week he was able to update them on the stage of development and answer any concerns in detail based on his analysis of the technical team's reflections. Once they completed the model, it was deployed by the department's IT team, and a protected balance sheet that recorded all the funds given to each registered small business was populated. The department was impressed by how efficient the intervention made the collaboration. They brought back Sean's team so they could create an intervention tool for all technical consultants in the department to use. Sean then provided the report to show the additional features and functionality he wanted to add to the intervention which further excited the department director.

4.4 Implementation Considerations

In addition to behavioral factors, the practical and technical logistics of interventions are critical for their success. This section delves into these implementation considerations for the design of behavior change interventions:

- *When is a suitable time to intervene*,
- *Where do the interventions take place*, and
- *How can we design effective interventions.*

To further illustrate the application of these dimensions, we present a usage scenario that demonstrates how these technical considerations can shape and facilitate the design of a responsible data science intervention tool in subsection 4.4.4

4.4.1 When: When is the suitable time to intervene?

The timing of behavior change interventions is a pivotal factor in their effectiveness[38]. Interventions ought to be strategically timed to align with key moments in the data

science process where they can have the most significant impact. Incorrect timing could render even the most well-designed interventions ineffective, as they may either preempt the need for action or come too late to influence the desired outcomes [133]. The timing of interventions can be informed by prior work in HCI on intervention and notification timing by Fogarty et al. [71]. Intervention developers can consider at least three different ways of characterizing "When" the intervention occurs: according to the **action occurrence**, the **phase in the data science process**, and the **stage of human information processing**.

Timing Based on Action Occurrence

The effectiveness of an intervention can greatly depend on its temporal relationship to the behavior it targets. Classifying interventions based on their timing relative to the behavior — whether they are **Synchronous** or **Asynchronous** — allows us to strategically influence data scientists' actions in a way that promotes ethical conduct and minimizes risk. Synchronous interventions are designed to work in real-time, providing immediate guidance or feedback during the occurrence of the behavior. Asynchronous interventions operate after the behavior has taken place, allowing for reflection and review.

Timing Based on Phase in the Data Science Process

The four stages of the data science process (preparation, analysis, deployment, and communication) as described by Crisan et al [47] are a sequence of interconnected stages, where each is crucial for the overall success of data-driven projects. Categorizing interventions according to these stages allows us to address the unique ethical and practical challenges that arise at each point:

1. **Preparation**: During the data preparation phase, interventions can be introduced to ensure data quality and integrity. For example, a Jupyter Notebook

plugin could automatically suggest privacy-preserving methods when sensitive data is being cleaned and prepared.

2. **Analysis**: In the analysis stage, real-time tools can assist data scientists by providing in-line guidance on statistical methods and algorithms that minimize bias and ensure fairness.
3. **Deployment**: During deployment, interventions can include mandatory ethical compliance checks that ensure models meet ethical standards before they are used in decision-making processes.
4. **Communication**: During the communication of results, interventions can help ensure that data visualizations and reports are transparent and do not mislead stakeholders about the implications of the data.

Timing Based on Four Stages of Human Information Processing

Another way to think about the timing of behavior change interventions is through the relevant stage of information processing. Human interactions and behaviors are guided by four steps of human information processing [142]: (1) information acquisition; (2) information analysis; (3) decision and action selection; and (4) action implementation:

1. **Information Acquisition**: Information acquisition refers to the acquisition and registration of multiple sources of information [142]. In the context of responsible data science, this stage involves gathering relevant data and information needed for analysis. It includes identifying sources, collecting and inspecting data, and ensuring its quality and relevance. For example, a Jupyter Notebook plugin could alert users when the data they are importing has historically been prone to bias or when the data lacks representation from certain groups. This plugin could provide links to additional resources or alternative datasets that might help balance or correct these biases.

2. **Information Analysis**: Information analysis involves conscious perception and manipulation of processed and retrieved information in working memory [14]. This stage also includes cognitive operations such as rehearsal, integration, and inference, but these operations occur prior to the point of the decision [142]. In the context of responsible data science, it includes applying statistical methods, algorithms, and models to understand the data. One potential intervention could be an embedded tool in Jupyter Notebook that analyzes the algorithms being used and suggests modifications or alternative algorithms that are known to reduce bias. This tool could also visualize the effects of bias in current models and offer real-time feedback on how changes to the model could improve fairness.

3. **Decision and Action Selection**: The Decision and action selection stage is where decisions are reached based on the iterations of the previous two cognitive processes [142]. Interventions at this stage help data scientists consider ethical implications and make informed, responsible decisions within the process of building a data science model. This involves supporting data scientists in making ethical decisions about which models to use or how to deploy them. For example, before finalizing a model, this system could ask questions to ensure the user has considered all ethical aspects, such as "Have you checked for gender bias in your model outcomes?" or "Does this model disproportionately affect a particular community?"

4. **Action Implementation**: Action implementation involves the implementation as a response or action consistent with the decision choice [142]. In the context of responsible data science, the final stage involves deploying models, sharing results, and ensuring that actions are carried out effectively. Automated tools could be integrated into Jupyter Notebooks to execute privacy-preserving techniques, such as data anonymization or differential privacy, automatically whenever data is exported or reports are generated. These tools could also im-

plement routine fairness checks before any analysis is finalized, ensuring that all outputs adhere to certain ethical standards.

4.4.2 Where: Where do the interventions take place?

The setting of a behavior change intervention profoundly influences its effectiveness [59]. The "Where" dimension analyses how seamlessly interventions integrate into the daily routines of data scientists, influencing both their usability and the likelihood of adoption. Properly situating interventions can bridge the gap between theoretical behavior change and practical, actionable modifications in real-world settings.

In this section, we describe two distinct approaches for embedding these interventions and their respective tradeoffs: **in-situ** and **ex-situ**. In-situ interventions are those that can be directly incorporated within the data science deployment environments (e.g., Jupyter Notebook [108], Google Colab, VSCode). By embedding behavior change prompts and guidance within the context of existing tooling, practitioners can receive real-time support during various stages of their workflow—from data preprocessing to model evaluation. On the other hand, ex-situ interventions exist outside of the tools used in data scientists' practices, extending their reach to standalone websites or systems like visual analytic platforms. For example, an ex-situ intervention might enable data scientists to export their project data to a dedicated ethical auditing tool outside their routine deployment environment.

4.4.3 How: How can we design effective interventions?

The "How" dimension addresses the various elements that must be considered to create interventions that are not only theoretically sound but also practical and engaging for the intended audience. This dimension fundamentally influences the usability, acceptance, and overall impact of the interventions. Some characteristics

of "How" to design effective interventions include **intrusiveness**, **programming requirement**, **interactivity**, and **level of automation**. These dimensions are not exhaustive but rather provide some exemplary considerations that can inform effective behavior change intervention design. The selection of these dimensions was informed by a combination of a literature review of existing frameworks in behavior change and human-computer interaction (HCI), along with iterative brainstorming sessions among the authors to ensure they capture the technical and practical needs specific to responsible data science interventions.

Intrusiveness

Intrusiveness concerns the visibility of interventions and the degree to which users can choose to engage with them. If an intervention is too intrusive, it may annoy users and lead to disuse; if too subtle, it might be ignored. Understanding the optimal level of intrusiveness helps in designing interventions that are effective yet respectful of the user's workflow:

1. **Hidden and Ignorable**: These interventions operate in the background with no notification to the user. For instance, an intervention in a Jupyter Notebook could silently monitor for the use of deprecated or non-compliant data processing methods, logging this use for later review without interrupting the user's workflow.
2. **Hidden and Not Ignorable:** These interventions operate in the background but take action without requiring user engagement, ensuring that essential tasks are performed. For example, an automated tool that corrects variable name errors without notifying the user. This type of intervention can improve the workflow by handling routine tasks silently but effectively.
3. **Visible and Ignorable**: These interventions are apparent but do not force interaction. An example could be a sidebar in Jupyter Notebook that displays

ethical guidelines or suggestions that users can choose to engage with or ignore at will during their work.

4. **Visible and Not Ignorable**: These interventions require user engagement to proceed. It should be utilized when a particular decision is critical. For instance, a popup that requires user action before certain types of data, such as sensitive or protected groups, can be processed.

Programming Requirement

The need for programming skills to utilize an intervention influences its accessibility and the breadth of its deployment. Interventions that **Require Coding** might limit their use to more technically adept users (e.g., a Jupyter Notebook extension could require users to implement custom scripts that check for bias in data before analysis can proceed), whereas those with **No Coding Required** can be adopted more widely across various levels of technical expertise (e.g., a pre-built Jupyter Notebook extension that automatically scans datasets for sensitive information and prompts users through a simple GUI to anonymize data before analysis).

Interactivity

The degree of interactivity in an intervention influences how engaging and adaptable it is. It could be categorized into two different types: **Interactive** and **Static**. Interactive interventions involve active participation or input from the user, such as tools that require users to make selections, provide feedback, or make decisions based on the provided information (e.g., an interactive module in Jupyter Notebook that simulates different data handling scenarios and asks users to choose the best ethical approach, providing instant feedback on their choices). On the other hand, static interventions do not allow for user input but provide prompts, information, notifications, or warnings (e.g., a static report generated by a tool within Jupyter Notebook

Levels of Automation		Definition
No Automation	1	The intervention offers no assisted decisions and actions, and human are fully responsible for them
Low Automation	2	The intervention offers a range of options but leaves the final decision to human
	3	The intervention uses predefined criteria to limit choices to the most appropriate ones.
	4	The intervention proposes the best action based on its analysis.
	5	The intervention executes a proposed action only after human confirmation.
	6	The intervention allows the human a restricted time to veto before automatic execution
High Automation	7	The intervention executes automatically, and only informs the human when necessary
	8	The intervention handles tasks independently and provides details only upon human's request
	9	The intervention informs the human only if it, the computer, decides to
Fully Automated	10	The intervention decides everything and acts autonomously, ignoring the human

Figure 4.4: We adopt the concept of levels of automation from Vagia et al. [171] to measure the intervention's automation level in the context of data science.

that assesses the ethical implications of a project's data usage, available for review at the user's discretion).

Level of Automation

The level of automation determines how much of the decision-making process is handled by the intervention versus the user. This balance is crucial as it affects the user's control over the tasks and their trust in the intervention's recommendations or actions. We adopt the concept of 10 levels of automation from Vagia et al. [171] and adapt it into the context of data science, as shown in Figure 4.4. These levels range from complete user control to full automation by the system.

For the sake of simplicity in our subsequent coding of existing responsible data science tools (Section 4.5), we group the total 10 levels of automation into four types as shown in the Figure 4.4: No Automation (level 1 in section 4.4.3), Low Automation(level 2-6 in section 4.4.3), High Automation (level 7-9 in section 4.4.3), and Fully Automated (level 10 in section 4.4.3).

4.4.4 Usage Scenario: A Professor's Intro to Responsible Data Science Course

Intervention Inception

Dr. Y is a computer science professor teaching a Fall course called "Introduction to Data Science." As Dr. Y was preparing the teaching plan for the summer, Dr. Y wanted to include a unit on responsible data science after covering basic data science concepts and skills. Dr. Y wants to conclude the responsible data science unit with a project in which the students execute responsible data science practices. To ground the project in the real world, Dr. Y chose to scope the project around creating a prediction model for loan approvals. Dr. Y selected the South German Credit dataset. The dataset includes credit and demographic information from clients with good and bad credit scores from 1973 to 1975 [2]. An important feature Dr. Y wants to focus on is the foreign worker feature. While Dr. Y is very excited about debuting the responsible data science project in the class, Dr. Y wants to ensure that the students engage in the current practices for addressing anti-immigrant bias. Dr. Y decides to build an intervention tool for the project. Dr. Y wants to encourage a reflexive development of responsible data science skills not a prescriptive development. Dr. Y wants to explore if in-situ explanation, guidance and reflection prompts students to change their behavior towards adopting responsible data science as a part of their everyday data science practice. Dr. Y refers to the Behavior Change for Responsible Data Science design space website to guide the design of the responsible data science intervention tool for the students.

Key Insights from Design Space

Dr. Y completed the Behavioral Considerations of the design space to build a comprehensive picture of the student user group based on previous iterations of the

course. Next, Dr. Y considers the Implementation Considerations. The **"When"** branch prompts Dr. Y to consider the finer points of the tool's design in terms of the data science lifecycle. Despite Dr. Y's interest in the different ideas, Dr. Y realizes the intervention would become too complex if it had to cover the majority of the data science lifecycle. Dr. Y decides to focus on the "analysis" stage which is the more ambiguous yet essential stage in practicing responsible data science. In the **"Where"** branch, Dr. Y decides that the tool should come in the form of an "in-situ" plugin for the coding notebook platform, Jupyter Notebook. Dr. Y's course only teaches students how to code in Jupyter Notebook so it's an environment the students are comfortable with. Finally, Dr. Y visits the **"How"** branch of the design space to decide the functionality of the tool. Dr. Y wants to encourage the student users to engage with the intervention before they can move forward. Pop-ups can be a feature that enforces this user experience ("visible and not ignorable"). Given the educational purposes of the intervention, Dr. Y doesn't want to create a complex and highly interactive tool. Therefore, the tool will be primarily "static" but allowing a drop-down to support students browsing results in different evaluation metrics ("No Automation").

Design Space Impact

After walking through the design space, Dr. Y downloads the consolidated behavior change for responsible data science report that contains all of their notes and selections. Dr. Y then uploads the report to their teaching plan folder and now feels more confident about completing the intervention tool over the summer before the course. Dr. Y refers to the report while writing the project requirements and development plan for the intervention tool. When two undergraduate students from a previous class express interest in working with Dr. Y over the summer, the report serves as one of the onboarding documents for their work over the summer. As Dr.

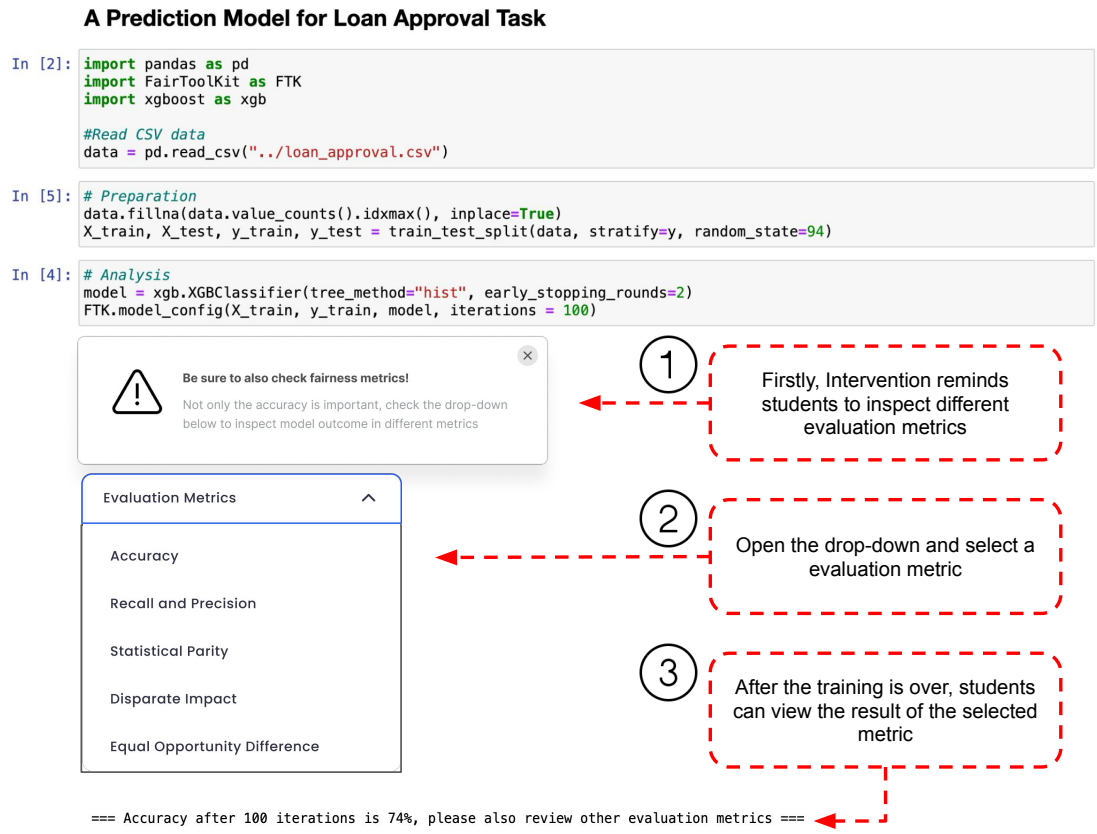

Figure 4.5: An exemplary intervention Dr. Y envisioned.

Y routinely meets with the research assistants to check on their progress, they all refer to the report to check if the team's progress aligns with the design imagined in the report. If changes need to be made, the team returns to the website to make new report iterations. When the prototype is deployed in Dr. Y's first Introduction to Data Science Class, Dr. Y shares the most recent report with the class as an act of transparency. As shown in Figure 4.5, the intervention first reminds students to inspect different evaluation metrics with a pop-up box. Furthermore, students can interact with the drop-down menu to measure the model's performance using different evaluation metrics. Once the training is over, students can view the result of the selected metric at the bottom of the drop-down menu. The intervention receives strongly positive feedback from the students so Dr. Y submits a manuscript to share the findings from their project and includes the report as a supplementary document

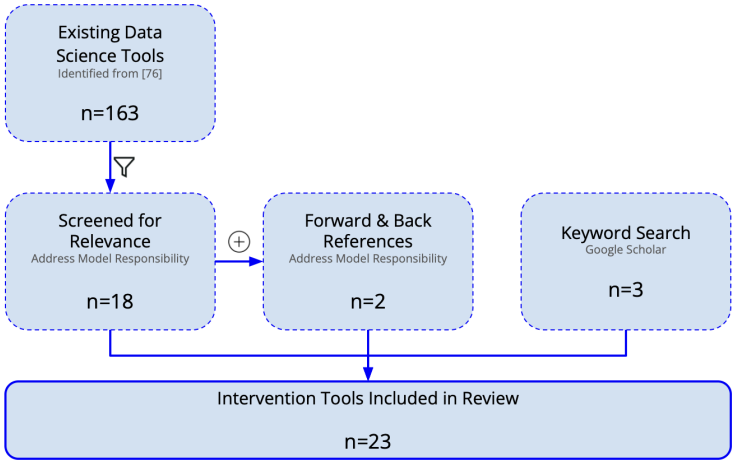

Figure 4.6: The process of arriving at the 23 interventions discussed in Section 4.5.

for readers to refer to. All in all, Dr. Y is glad they took the time to work through the design space because it improved the productivity of the project, kept collaborators on the same page, and provided a method of transparency for users.

4.5 Characterizing Existing Intervention Tools

To demonstrate the utility and applicability of our proposed design space, we conducted a targeted survey and coding of existing tools in the domain of RDS. The objective of this analysis is twofold: first, to map the features of these tools to the implementation considerations (When, Where, and How) of the 5W1H dimensions to understand coverage of the design space; and second, to identify trends, gaps, and opportunities for further innovation in RDS. For this analysis, we do not report on the behavioral dimensions (Why, Who, and What), since this would require us to make inferences about the developers' intentions for the tools, which is not always explicit for these artifacts.

4.5.1 Method

To identify relevant behavior change intervention tools for RDS, we began by reviewing the survey conducted by Wang et al. [184] which covers 163 existing tools that facilitate data science practices (Figure 4.6). From this set of 163 existing data science tools, we assessed their relevance by reviewing available abstracts, full papers, GitHub repositories, prototypes, and demo videos, where applicable. We specifically selected 18 tools that directly addressed issues related to model responsibility (e.g., What-If-Tool [189]) or ethical considerations (e.g., DocML [21]) in data science. Following this initial filtering, we conducted forward and backward literature searches as well as keyword searches on Google Scholar to identify additional relevant intervention tools that focus on RDS in the past 10 years. This involved reviewing papers cited and cited by the filtered tools, further expanding our dataset of tools for RDS.

In total, 23 RDS intervention tools were included as they either: (1) directly supported responsible model deployment practices, such as subgroup analysis [193], bias auditing and reduction [152, 136, 189], model outcome evaluation and monitoring [6, 132, 180, 95], fair model building [113, 182], effective communication through sensemaking visualizations [80, 112]; or (2) contributed to ethical considerations more broadly, such as by providing machine learning documentation for ethical priming during model deployment [21, 84, 196], highlighting the consequences of configuration changes on fairness [110, 186, 185], or explaining model interpretability [131, 138, 183, 188].

Next, three authors collaboratively developed the codebook through an iterative process. The team held four working sessions, during which we discussed each dimension of the design space and how it would apply to the coding process. During these sessions, we refined the definitions of each subcategory to ensure they accurately reflected the features of the tools being assessed. Each session resulted in revisions to the codebook, which was then piloted on a small set of tools to ensure consistency

in interpretation and application. This codebook is attached in the supplementary materials. It served as the guiding framework for coding each tool based on the technical considerations (When, Where, How) outlined in the design space. Two authors then independently coded the tools across four rounds of coding, and 5-7 uncoded tools were coded in each round. After each round, the coders assessed their mutual agreement using Cohen's Kappa [121] to measure inter-rater reliability. Inter-rater reliability in each of the four rounds of coding were $\kappa_1 = 0.31$, $\kappa_2 = 0.39$, $\kappa_3 = 0.72$, and $\kappa_4 = 0.89$, respectively, for an overall inter-rater reliability of 0.62. This process ensured consistent application of the codebook definitions and allowed for iterative refinement of the coding process. At the conclusion of each round, the coders reviewed their results to reach a consensus coding, resolving discrepancies, and refining the codebook as necessary.

4.5.2 Results

Below, we present a summary of the coding results, focusing on the technical dimensions of When, Where, and How. **F1-F8** describe 8 salient findings. Figure 4.7 provides a visual representation of these results, with key findings highlighted below:

F1: *(How)* Minimal automation to develop responsible skillset across tools. As described in section 4.4.3 and Figure 4.4, we categorized interventions into four levels of automation: No Automation (level 0), Low Automation(level 1), High Automation (level 2), and Fully Automated (level 3). The majority of tools either offer no automation (52%) or a low level of automation (35%), with only 13% of tools offering high levels of automation. This suggests that developers may prioritize maintaining human agency in RDS, likely due to the complex ethical judgments involved, which may not be easily navigated by fully automated systems.

Low-automation interventions leave all decision-making and behavior choices to

		When										Where		How									
		Action Occurence		Data Science Process				Stages of Human Information Processing						Intrusiveness				Programing Requirement			Interactivity		
Year	Intervention Name	Synchronous	Asynchronous	Preparation	Analysis	Deployment	Communication	Information Acquisition	Information Analysis	Decision and Action Selection	Action Implementation	In-Situ	Ex-Situ	Hidden and Ignorable	Hidden and Not Ignorable	Visible and Ignorable	Visible and Not Ignorable	Coding Needed	Coding not Needed	Level of Automation	Interactive GUI	Static Information	
2019	Aequitas		☑	☑	☑	☑				☑		☑	☑			☑		☑		0	☑		
2019	VizSeq		☑			☑			☑			☑	☑			☑		☑		1	☑		
2019	InterpretML	☑	☑		☑	☑		☑	☑			☑				☑		☑		1	☑		
2019	Interpret-Community	☑			☑	☑			☑			☑				☑		☑		0	☑		
2019	What-if Tool		☑		☑				☑	☑		☑				☑			☑	0	☑		
2020	Whatlies		☑	☑	☑			☑	☑	☑		☑				☑		☑		0	☑		
2020	RAI Widgets	☑			☑	☑				☑		☑				☑		☑		0	☑		
2022	TimberTrek		☑		☑			☑	☑	☑		☑	☑			☑			☑	1	☑		
2022	GAM Changer		☑		☑	☑				☑		☑	☑			☑			☑	0	☑		
2022	Evidently		☑		☑	☑			☑			☑	☑			☑		☑		0	☑		
2022	Visual Auditor		☑			☑		☑	☑			☑	☑			☑			☑	2	☑		
2023	CausalVis	☑			☑				☑	☑		☑				☑		☑		0	☑		
2023	Calibrate		☑			☑			☑			☑				☑		☑		0	☑		
2023	DocML	☑		☑	☑	☑				☑		☑				☑	☑		☑	0	☑		
2023	EDAssistant	☑		☑				☑	☑	☑		☑				☑		☑		2	☑		
2023	ModelSketchBook		☑		☑					☑		☑				☑		☑		1	☑		
2023	Notable		☑				☑			☑	☑	☑				☑		☑		1	☑		
2023	VizProg	☑		☑				☑	☑				☑			☑			☑	1	☑		
2023	watsonx.governance	☑			☑	☑		☑	☑	☑			☑			☑	☑		☑	1	☑		
2024	HAX Toolkit	☑		☑	☑					☑			☑			☑			☑	0	☑		
2024	Farsight	☑			☑			☑	☑	☑		☑				☑			☑	1	☑		
2024	Wordflow	☑			☑	☑		☑	☑	☑			☑			☑			☑	0	☑		
2024	Retrograde	☑		☑	☑			☑	☑	☑		☑				☑			☑	2	☑		
Percentile		52%	52%	30%	74%	52%	4%	43%	70%	70%	4%	83%	43%	0%	0%	100%	9%	52%	48%	12:8:3:0	100%	0%	

Figure 4.7: Summary of coding results for behavior change intervention tools in RDS.

the data scientist (e.g., DocML [21] only reminds users to follow the model cards proposal during model development). In contrast, high-automation interventions handle most initial decisions, involving users only for confirmation or when necessary (e.g., EDAssitant [113] automatically searches and recommends relevant Python APIs and notebook examples, asking users to confirm their selection). High levels of automation, while efficient, may not yet be trusted to navigate these complexities without risking unintended biases or oversights. However, it is still difficult for data scientists to navigate complex ethical considerations even with interventions [49].

Takeaway: As this research area grows, limiting automation in intervention tools may be viewed as an essential feature. Rather than automating the responsible work, interventions can be designed to illustrate and teach RDS practices through guided actions for users. Therefore, one potential reason behind this observation is developers are likely to prioritize manual or semi-automated tools to ensure human oversight and control, preserving ethical accountability in decision-making processes.

F2: *(How)* Interventions are visible, but ignorable. All coded tools are visible but ignorable (100%), reflecting a design preference towards non-intrusive interventions. This method could be the result of balancing usability with ethical guidance by not disrupting user control.while not engaging users too frequently. However, this also highlights a potential area for improvement, as critical ethical considerations may sometimes require more visible and not ignorable interventions, especially in high-risk scenarios in which enforcing ethical behavior is essential (e.g., interventions that facilitate building crime recidivism prediction model). This dimension also highlights how RDS intervention tools consider user agency. As outlined in the "What" branch of our design space, internalization is the strongest avenue for attitude change (see section 4.3.3). Choosing to execute ignorable suggestions over time can encourage users to adopt RDS practices on their terms. On the other hand, in the case of high-risk domains, a compliance approach to attitude change may be preferred for its expediency. Another possible reason why we coded all interventions as "visible and ignorable" could be the bias in our search process for data science tools.

Takeaway: We encourage the development of non-ignorable interventions, especially for high-stakes analysis scenarios.

F3: *(How)* Interactive interfaces dominate but there should be consideration of cognitive load. All coded tools provide an interactive GUI (100%), with none relying solely on static information. This could suggest the need for user engagement in RDS. One potential reason is ethical decision-making often requires dynamic feedback and user exploration to address evolving challenges effectively. Incorporating static information alone may not provide the flexibility or depth required to address the evolving nature of ethical challenges in data science workflows. Conversely, although static information alone may lack flexibility or depth, it presents an opportunity to reduce cognitive load for users. Static interventions can simplify decision-making by offering clear, concise guidance without overwhelming users with

too many options or interactions[145]. This reduced complexity could be beneficial in scenarios where quick ethical checks are needed, or when practitioners are already managing high cognitive demands from other tasks.

Takeaway: Intervention designers should strike a balance between dynamic user engagement and concise presentation of information to avoid cognitive overload.

F4: *(How)* Customization tradeoffs of coding requirements in interventions Some tools required coding, others not. Half of the tools (52%) do not require coding, indicating accessibility of RDS interventions to practitioners with varying levels of technical expertise. This trend aligns with efforts to democratize responsible data practices across different user groups [118]. No-code tools allow users to focus on developing code for the task or project at hand. However, there is a tradeoff between the ease of use offered by no-coding tools and the customization that coding tools provide. Tools that require coding allow users to tailor interventions more precisely to their specific needs, while no-coding tools prioritize simplicity and accessibility but may sacrifice customization.

Takeaway: In addition to no-code base functionality, intervention designers should consider providing the option to execute code within their tool for users to further customize intervention actions to their context.

F5: *(Where)* Preference for in-situ over ex-situ tools for accessibility. 19 RDS intervention tools (83%) are designed as in-situ tools. These intervention tools are integrated within the working environments of data scientists as notebook plugins or compatible Python packages. 26% of tools support both in-situ (within notebook) and ex-situ formats (standalone websites or toolkits). This emphasis on in-situ design could suggest the need for tools to be readily accessible and seamlessly embedded within existing workflows. Designers of data science tools often prioritize seamless integration in the workflow based on user feedback [193].

Takeaway: In-situ intervention designs can prioritize ease of use. Limiting barriers

to use provides ample opportunity for engaging in RDS practices.

F6: *(When)* Opportunities to intervene at later stages of lifecycleConcentration on information analysis and decision stages. Most intervention tools focus on the Information Analysis (70%) and Decision and Action Selection (70%) stages, with 36% of tools supporting both stages simultaneously. Interventions focusing on Information Analysis help data scientists process and interpret data ethically by providing insights into potential biases or fairness issues within the data. These interventions ensure that ethical considerations are embedded in the analysis process, and assist users in making responsible decisions during model building. This suggests that interventions prioritize assistance in data interpretation and decision-making, with fewer tools addressing the other stages; only 43% of interventions support Information Acquisition (e.g., TimberTrek [183] helps users to summarize different levels of the decision tree model at scale) and 4% of interventions support Action Implementation (e.g., Notable [112] supports users converting data findings into visualization story-telling). This suggests a potential gap that future interventions could concentrate more on either the Information Acquisition stage or the Action Implementation stage. For example, interventions could help data scientists gain deeper insights of potential correlations within data, or help users run fairness examination of model outcomes once they finalize model configurations.

Takeaway: While supporting initial data analysis is paramount for RDS, intervention designers should also explore how to conduct fairness evaluation and tuning at the later stages.

F7: *(When)* Emphasis on the analysis phase of data science. Most tools target the Analysis stage (74%), followed by the Deployment (52%) and Preparation (30%) stages. A notable gap exists in the Communication phase, where only 1 intervention tool (4%) provides support [112]. This suggests a potential area for future tools to enhance ethical communication and reporting of data science results.

For example, interventions could help standardize ethical reporting practices across projects, providing templates or prompts to ensure that all relevant ethical factors are included in final reports and visualizations. There is a lack of cohesive support across all stages of data science workflows. Existing tools tend to focus on isolated aspects rather than providing end-to-end support. Such fragmentation and unevenness reflect the complexity and dynamic nature of ethical challenges in data science. Different stages of the workflow involve varying stakeholder priorities and levels of urgency, making it difficult for existing tools to address RDS holistically.

Takeaway: Given the prevalence of intervention tools for analysis, future designers can address the lack of RDS support in the other stages of the life cycle (especially the communication stage).

F8: *(When)* Balance between synchronous and asynchronous. The distribution between synchronous (52%) and asynchronous (52%) interventions is evenly distributed (one intervention supports both synchronous and asynchronous [138]). This balance highlights the importance of addressing ethical concerns both in the moment, when critical decisions are made, and after the fact, when there is time for deeper consideration of long-term impacts. Recently, RDS scholarship is increasingly embracing reflexive techniques to contend with the complex decisions practitioners have to make [90]. Interventions can play an important role in spurring reflexive practices as a consciousness-raiser or potential collaborator in a user's RDS journey.

Takeaway: The presence of both intervention types suggests ethical data science workflows can benefit from both immediate guidance and opportunities for reflective evaluation.

4.6 Discussion

Generalizability and Robustness: While this framework was validated with a spe-

cific set of data science tools, its guiding principles—behavioral and implementation considerations—can be broadly applied to different contexts beyond those explored in this study. For example, developers creating intervention tools for domains such as medical data science can leverage the framework by considering e.g., the Regulations (Why) and Level of Automation (How) that is relevant and standard practice in this field. Furthermore, the separation of behavioral and implementation factors allows for incremental adoption in the given context; practitioners can prioritize dimensions that align with their immediate goals while gradually expanding their interventions to include more comprehensive support. The modularity of the design space can also be iteratively refined to enable users to adapt interventions as technologies, regulations, and societal expectations evolve.

4.7 Limitations:

While the 5W1H interrogative framework provides a scaffold for designing behavior change interventions in responsible data science, several limitations warrant acknowledgment. First, our design space may oversimplify the multifaceted nature of ethical decision-making in data science practices. By structuring interventions around discrete dimensions, we risk fragmenting what is inherently an integrated, context-dependent process. Ethical considerations in data science rarely fit neatly into categorizations, and the interactions between dimensions (e.g., how the "When" influences the effectiveness of "How") remain underexplored in our framework. Future work should explore these complex interdependencies to ensure interventions address the holistic nature of ethical decision-making rather than isolated components.

Second, although we have established a design space in this chapter to answer **RQ2**, we have yet to establish an evaluation plan to validate the efficacy of the

interventions that are created following our design space guidelines. Therefore, in chapter 5 and chapter 6, we aim to implement an intervention following our design space framework and explore methodologies for evaluating the efficacy of these interventions. This approach will help bridge the gap between theoretical design and practical impact assessment, furthering our thesis goal of developing and evaluating behavior change interventions that promote responsible data science.

Finally, our design space does not yet provide definitive guidance on how to resolve potential conflicts between different intervention strategies or how to prioritize among multiple design considerations. As we move toward the practical implementation and evaluation of interventions in the subsequent chapters, addressing these trade-offs will become increasingly important for translating our theoretical and design contributions into meaningful improvements in responsible data science practice.

4.8 Summary

In this project, we conducted work to address **RQ 2** (*How can I scaffold the design and development of behavior change interventions for responsible data science?*) by exploring the essential role of behavior change interventions in advancing responsible data science practices. Addressing the complex ethical challenges in data science, we aim to foster ethical decision-making and responsible model deployment through a multifaceted approach that combines technical skills, ethical awareness, and behavioral insights. We aim to catalyze a cultural shift towards ethical data practices within the data science community. To achieve this, the project outlines a design space for behavior change interventions, guided by the 5W1H framework (Who, What, When, Where, Why, and How). This framework helps in identifying the target audience, desired behaviors, optimal timing, location, objectives, and methods of interventions. We examined 23 existing responsible data science tools and mapped their functionali-

ties to our design space, identifying gaps and potential opportunities for future work. Additionally, we demonstrated the usability of this design space through two usage scenarios to show how it can be applied at the ideation phase for building effective tools to foster responsible data science practices.

Chapter 5

Developing a Behavior Change Intervention for Technical Responsibility in Data Science Pre-Processing

5.1 Motivation

Building upon the behavior change framework for responsible data science established in chapter 3 and chapter 4, this chapter moves from theoretical foundations to practical implementation. New techniques have been developed to tackle challenging and complicated tasks, many of which require human's input due to potentially life-changing ramifications if models are not fair (e.g., hiring[119], loan approval[168], and crime recidivism prediction[69]). One notable example of AI bias is the COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) algorithm used in US court systems to predict the likelihood that a defendant would become a recidivist. The model predicted twice as many false positives for recidivism for black

offenders (45%) compared to white offenders (23%)[69]. Thus, while data science and machine learning models are typically trained to maximize prediction accuracy and model robustness, it is apparent that social fairness and parity are important objectives to manage during model training and performance evaluation. These examples highlight the systemic inequities that data science pipelines can perpetuate when left unchecked—precisely the problem identified in chapter 1.

While some technical tools that promote responsible behaviors in DS exist to measure and mitigate bias in the model pre-processing, in-processing, and post-processing stages of model development, an often minimized part of the DS pipeline is the potential downstream effect on fairness when different pre-processing strategy sequences are applied to data. Not only does the prevailing literature emphasize the critical importance of the pre-processing phase in ensuring fairness within DS and ML, but the influence of pre-processing is corroborated by our experimental findings (section 5.2).

After establishing a conceptual framework of responsible data science using behavior change theories in chapter 3 to address **RQ1**, we seek to develop a technically responsible behavior change intervention for responsible data science. In this project, **we introduce a visual analytic prototype, PreFair, designed to enhance technical responsibility in DS and ML by assisting model builders in exploring the trade-offs between various pre-processing strategies.** In the context of the design space from chapter 4, PreFair exemplifies a targeted intervention that enhances Capability and Opportunity within the COM-B model (What) for data scientists and students (Who). This ex-situ tool (Where) focuses on the pre-processing stage (When), using an interactive, programming-free interface (How) to promote responsible practices and demonstrate the connection between pre-processing decisions and model fairness (Why). PreFair allows for the evaluation of model performance using both confusion matrix and fairness metrics, considering different sequences of pre-processing techniques. Through this exploration, we aim to shed

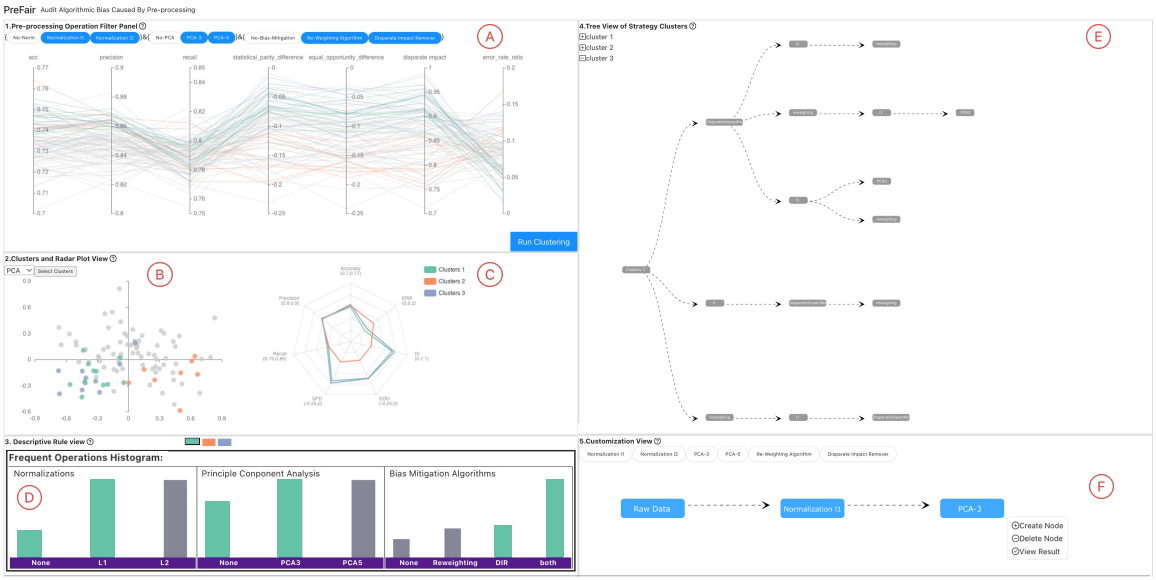

Figure 5.1: PreFair integrates multiple views for assisting users in choosing a fairness-aware pre-processing strategy, including: **A.** Parallel Coordinates View shows pre-processing strategies alongside different fairness metrics; **B.** Cluster View shows pre-processing strategies based on similar fairness outcomes; **C.** Radar Plot View shows strategies' performance under various metrics; **D.** Rule View characterizes how strategies in the selected cluster are different from others in the format of descriptive rule lists; **E.** Tree View visualizes the strategies of each cluster in the form of a word-tree structure; **F.** Customization View provides a playground for users to test their own pre-processing strategies.

light on the pivotal role of technically responsible behavior change interventions designed for data pre-processing and its relative importance within the broader context of DS and ML fairness. This research project addresses the goal of development in **RQ3**: *How effective are behavior change interventions at improving responsible data science outcomes? (Goal III: Develop and Evaluate Interventions).* Currently, this research project[54] is in preparation for submission to TVCG 2026.

5.2 Quantifying the Impact of Pre-Processing on Model Fairness

In this section, we present a machine learning experiment in which diverse pre-processing strategies are applied to the same dataset and model. This demonstration serves to emphasize the substantial influence that pre-processing can have on model performance, especially on model fairness. We first introduce (i) the specific pre-processing operations, (ii) privileged/unprivileged groups, and (iii) evaluation metrics utilized in both this experiment and PreFair. Subsequently, we delve into (iv) the experiment and its outcomes, showcasing the significant impact that diverse pre-processing strategies can have on model performance and fairness.

In the context of the forthcoming experiment and the PreFair prototype, we utilize the German Credit Dataset, an Adaboost classifier, six pre-processing operations, and seven evaluation metrics to illustrate the system and contextualize experimental findings. It is worth noting, however, that the experimental findings and system techniques we present are not limited to these choices and can be replaced with alternative datasets, models, pre-processing operations, and fairness metrics as appropriate. We discuss the generalizability of PreFair to cover more machine learning phases and options in the discussion section (chapter 7).

Pre-Processing Operations. We incorporate 6 canonical frequently used pre-processing operations for the experiment and proof-of-concept prototype in PreFair. These operations fall into three categories: data normalization, dimension reduction, and bias mitigation operations. Users can pick Manhattan Distance (L1) or Euclidean Distance (L2) to *normalize* the dataset, apply Principle Component Analysis (PCA) with 3 or 5 components (PCA-3 and PCA-5, respectively) to *reduce the data dimensionality*, and utilize the Re-weighing Algorithm[102] or Disparate Impact Removal Algorithm[74] to *reduce bias* towards the unprivileged group before feeding the dataset

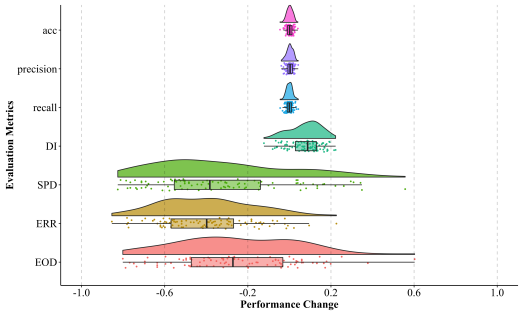

Figure 5.2: The percentage change in model metrics for the German credit dataset[58] after applying different pre-processing strategies. A positive or negative percentage change refers to an increase or decrease in the performance of respective evaluation metrics, compared to training the model without any pre-processing. DI, SPD, ERR, and EOD are the abbreviations for disparate impact, statistical parity difference, error rate ratio and equal opportunity difference, respectively.

into the model for training. Multiple impractical strategies are excluded from the option space (e.g., normalizing the dataset twice with L1 and then L2 Normalization will not make sense).

Privileged/Unprivileged Groups. In the German Credit Dataset[58], as discussed in AI Fairness 360[18], young (age<25) loan applicants are shown to be statistically less likely to get loan approval compared to older (age≥25) applicants, and female applicants are similarly observed to have lower approval rates compared to male applicants. Therefore, we define the unprivileged group as young (age<25) female applicants and the privileged group as older (age≥25) male applicants in this project.

Evaluation Metrics. We chose to use a combination of three performance metrics[162]: Accuracy, Precision, Recall; and four fairness metrics: Statistical Parity Difference[40], Equal Opportunity Difference[82], Disparate Impact[159], and Error Rate Ratio to quantify pre-processing strategies' predictive performance. Figure 5.3 describes the performance metrics used in PreFair, where $\hat{y}$ represents an instance's predicted label, **P** represents the privileged group, and **UP** represents the unprivileged group.

Performance Metrics	Formula	Description
Accuracy	$\frac{TP+TN}{P+N}$	ratio of correctly predicted instances to entire dataset
Precision	$\frac{TP}{TP+FP}$	ratio of *true positives* over all predicted positives.
Recall	$\frac{TP}{TP+FN}$	measure of the model correctly identifying *true positives* over all actual positives
Statistical Parity Difference	$Pr(\hat{y}=1\|Group=UP) - Pr(\hat{y}=1\|Group=P)$	difference in probability of positive prediction between protected (*unprivileged*) and *privileged* groups
Equal Opportunity Difference	$\left(\frac{TP}{TP+FN}\right)_{Group=UP} - \left(\frac{TP}{TP+FN}\right)_{Group=P}$	difference in true positive rates between the protected (*unprivileged*) and *privileged* groups
Disparate Impact	$\frac{Pr(\hat{y}=1\|Group=UP)}{Pr(\hat{y}=1\|Group=P)}$	ratio difference between positive prediction rate for the protected (*unprivileged*) vs. *privileged* groups
Error Rate Ratio	$\frac{\left(\frac{FP+FN}{P+N}\right)_{Group=UP}}{\left(\frac{FP+FN}{P+N}\right)_{Group=P}}$	ratio difference between prediction error rates for the protected (*unprivileged*) vs. *privileged* groups

Figure 5.3: TP, TN, FP, and FN are abbreviations for *true positive*, *true negative*, *false positive* and *false negative*. These are the possible outcomes of making predictions with a binary classifier. N and P are the total *negative* and *positive* cases respectively. P and UP represent the *privileged* and *unprivileged* groups respectively. $\hat{y}$ is the predicted value; thus $Pr(\hat{y} = 1)$ indicates the probability of a positive prediction.

Experiment and Results. We conducted a comprehensive experiment to explore the impact of different pre-processing strategies on model performance using the German Credit Dataset. All possible permutations of the six pre-processing operations were applied to the dataset. The experiment employed an Adaboost Classifier with 100 estimators to predict loan approval for individual candidates, with a 7:3 train/test split ratio. We utilized a stratified 10-fold validation from Scikit-learn to calculate average model performance under the confusion matrix and four fairness metrics.

The empirical results in Figure 5.2 illustrate the percentage change in model metrics when various pre-processing strategies were applied before model training. Positive and negative percentage changes indicate increases or decreases in the performance of evaluation metrics compared to training the model without pre-processing. Notably, while different pre-processing strategies had minimal impact on the confusion matrix, they notably influenced the four selected fairness metrics. This observation aligns with the findings reported by Biswas and Rajan [24], highlighting the important role the pre-processing stage plays in machine learning fairness.

5.3 Design Approach

In this section, we describe our design approach for PreFair including our design process and design goals.

5.3.1 Design Process

PreFair is the result of an iterative design process. We implemented our initial design after more than a dozen rounds of sketches and discussions among the four authors. The initial prototype was designed around the clustering of pre-processing strategies. We then invited three VIS+HCI researchers and three ML experts to provide formative feedback on the usability and the interaction design of the prototype. Formative feedback was primarily focused on supporting a more cascaded and user-friendly workflow to narrow down the search space. Some specific feedback included (i) supporting multiple clustering algorithms, (ii) removing unnecessary interactivity from rule learning and using it instead in a purely descriptive fashion, (iii) supporting additional linking between several views, and (iv) enabling comparison of different strategies' performance in various ways. We incorporated this feedback in an updated prototype design, described in Section 5.4, before collecting summative feedback from six additional experts (who did not participate in the formative feedback phase).

5.3.2 Design Goals

Based on our iterative design process, we synthesized the following three design goals (**DGs**) to guide the development for PreFair:

DG 1. Present strategies' outcomes under various evaluation metrics. Data pre-processing choices can lead to substantially different outcomes according to different evaluation metrics. Overemphasizing specific metrics or a specific pre-processing strategy could lead to blind spots in the model deployment, such as a lack

of awareness of alternative pre-processing strategies which could have even better fairness outcomes. This objective translates to presenting a model's predictive results given all possible pre-processing strategies under a variety of evaluation metrics.

DG 2. Enable nuanced strategy exploration and comparison. Deciding on a proper pre-processing strategy demands abundant knowledge of the option space and subsequent exploration and comparison. Therefore, we aim to assist users to explore the nuances of the option space with user-friendly approaches to create, evaluate, and compare strategies.

DG 3. Facilitate narrowing down the option space. In addition to facilitating the exploration of the option space, we also aim to help users *narrow down* the search space to select a pre-processing strategy that works for their needs. Because it is unlikely to satisfy all fairness criteria simultaneously [101], we aim to capture and present trade-offs of strategies based on differences in evaluation metrics and support a step-by-step filtering process so users can select an appropriate pre-processing strategy.

5.4 Visual Analytic Interface

The PreFair user interface consists of six views, as shown in Figure 5.1. **Parallel Coordinates View** (Figure 5.1A) visualizes each pre-processing strategy as a polyline with each vertical axis representing the strategy's performance for an evaluation metric. **Cluster View** (Figure 5.1B) clusters pre-processing strategies together based on similarity of evaluation metric outcomes. **Radar View** (Figure 5.1C) compares clusters and designated strategies across different metrics. **Rule View** (Figure 5.1D) characterizes the relationship between strategies inside and outside the selected cluster. **Tree View** (Figure 5.1E) presents an organized representation of strategies in each cluster using a word tree representation where common sub-sequences are

collapsed into a single node. **Customization View** (Figure 5.1F) supports brainstorming, customizing, and comparing specific strategies' outcomes. We describe these views in greater detail in the following three sections, which functionally correspond to providing an overview of strategies (Section 5.4.1), narrowing down the search space (Section 5.4.2), and comparing specific strategies (Section 5.4.3).

In Figure 5.4, we illustrate the workflow of PreFair and show how these six views interact with one another to facilitate users exploring trade-offs and narrowing down the space of different sequences of pre-processing strategies under the evaluation of both confusion and fairness metrics. Initially, **A**, all strategies are screened using specific inclusion criteria pertinent to pre-processing operations. Following this, **B**, the Cluster View, enables further filtering by grouping strategies based on their similarities in evaluation metric outcomes. If the resulting clusters are deemed insufficient, users have the flexibility to **C**, revise the filters with assistance from the **D**, Rule View. Alternatively, they may **E**, opt for a different clustering method. Once a satisfactory clustering arrangement is achieved, **F**, users can undertake a comparative analysis of the retained strategies through four integrated views to finalize their choice. Throughout this process, the option to **G**, return to the Parallel Coordinate View is available, allowing users to modify the initial filters based on any new insights or considerations that may arise.

Because there are inevitably more strategies than can be plausibly considered simultaneously by a human, we use clustering to automatically discover groupings of similarly performing strategies. We chose complementary techniques with the Parallel Coordinates and Cluster Views to make exploration of these strategies productive. The scatter plot in the Cluster View makes it easier to spot groupings that emerge based on similar evaluation metrics, while the Parallel Coordinates View makes it possible to see how various strategies or groups of strategies perform with respect to individual metrics. The clusters are mapped to color in both the Parallel Coordinates

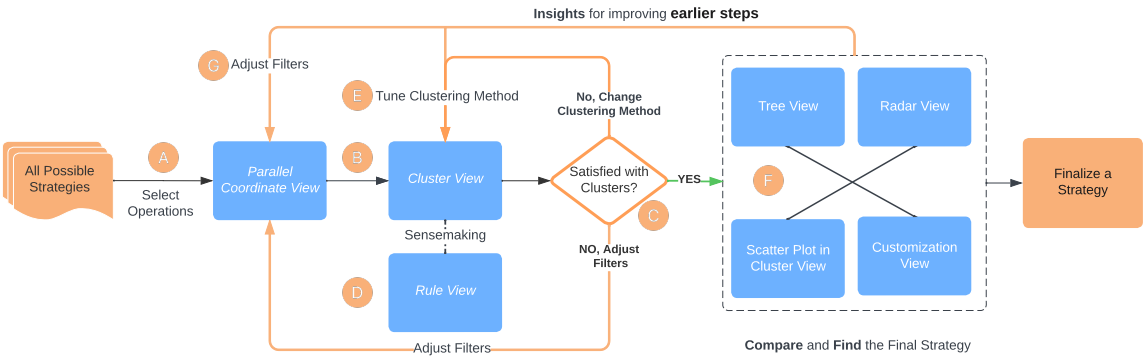

Figure 5.4: The workflow for PreFair. All strategies are **A.** filtered based on inclusion criteria for pre-processing operations. Next, the **B.** Cluster View facilitates additional filtering based on strategies that result in similar evaluation metric outcomes. If the clusters are not sufficient, users can **C.** adjust the filters again with the help of the **D.** Rule View, or **E.** change the clustering method. Once satisfied with the clustering strategy, **F.** users can compare the remaining strategies using four coordinated views to finalize their choice. During this process, users could **G.** go back to the Parallel Coordinate View to adjust the filters if they come across insights to improve the initial filtering process.

and Cluster Views so individual strategies can be evaluated, but we also provide a Radar View to give an overview of the clusters.

5.4.1 Overview of Strategies

Users will encounter the prototype first through an overview of all strategies' performance across all evaluation metrics, providing a "big picture" view of the distribution of strategies and metrics. Because it is designed to support comparison and filtering, the Parallel Coordinates View supports **DG 1** and **DG 2** by explicitly visualizing strategies' performance and distribution.

Parallel Coordinates View: Above the Parallel Coordinates View, users can choose which pre-processing operations to include or exclude by toggling the operation buttons, and the Parallel Coordinates View will dynamically update to show all possible pre-processing operations which contain at least one of the chosen operations. As shown in (Figure 5.1A), the parallel coordinates view represents each

pre-processing strategy as a poly-line in the plot, where the intersection with each vertical axis indicates the evaluation metric for the given strategy.

With a large number of possible pre-processing permutations, the Parallel Coordinates View can suffer from occlusion and excessive edge crossings. To mitigate this issue, our design focuses on enabling effective filtering and clustering to reduce the visual complexity. Users can interact with the Parallel Coordinates View by clicking and dragging along each axis to create filters for an acceptable range of values for each evaluation metric. Users may choose to perform initial filtering using this approach to get rid of strategies whose performance according to certain metrics would be unacceptable (e.g., users may apply a filter to only include options with a minimum of 75% accuracy). By allowing users to filter out less relevant strategies early on, we help minimize the number of lines displayed simultaneously.

Next, clicking on the "Run Clustering" button will group the remaining candidate strategies into multiple clusters for further exploration by a clustering method that comes with the best silhouette score (a measurement of cluster quality, explained further in the next section). Because the clustering approach groups similar strategies together, it can further highlight patterns within clusters, which simplifies the analysis (described in further detail in subsection 5.4.2). This combined filtering and clustering strategy helps maintain the comprehensibility of the Parallel Coordinates View even when handling complex datasets.

5.4.2 Narrow Down the Search Space and Explain Options

PreFair is designed to support the process of narrowing down the search space of pre-processing strategies. To accomplish this goal (**DG 3**), we employ clustering and descriptive rule learning. *First* we apply clustering to separate significantly differentiated strategies into groups whose member strategies have similar evaluation metric outcomes. *Second* we present the centroid strategy of each cluster in a radar

plot showing the performance of each cluster along different metrics represented as the radial spokes. We incorporate descriptive rules to characterize the distinctions between clusters. *Lastly*, users can select a cluster that represents the ideal balance of evaluation metrics. Alternatively, users can click on the "select clusters" button to switch to another clustering method, or revisit the Parallel Coordinates View for further filtering.

Cluster View: After the initial round of filtering in the Parallel Coordinates View, we cluster and visualize the qualified strategies in the Cluster View. Clustering the strategies can narrow down the selection of strategies by identifying pre-processing sequences which result in comparable performance and fairness outcomes. The high-dimensional strategy performance evaluated by 7 metrics (3 performance metrics and 4 bias metrics) is clustered and projected into a 2-dimensional scatter plot in the Cluster View. Each point shown on the scatter plot represents a pre-processing strategy, and the color encodes the cluster to which the strategy belongs. The scatter plot provides users with a visual overview of how different strategies group together based on their performance across multiple metrics. Users can thus quickly identify groups of strategies that perform similarly, allowing them to focus on specific clusters rather than individual strategies. This helps to significantly narrow down the search space.

The scatter plot in the Cluster View also guides users by showing the spatial relationship and density of different strategy clusters. For instance, if a cluster appears densely packed, it indicates that the strategies within this cluster perform similarly across the selected metrics. Conversely, more sparse clusters might suggest more variability in performance, prompting further investigation. By selecting a strategy from the scatter plot, users can leverage the aggregated insights provided by the clustering process. This is particularly useful when the Parallel Coordinates View

reveals too many individual strategies, making it difficult to identify patterns. In contrast, the scatter plot's clusters provide a higher-level overview that simplifies decision-making.

Users can swap the clustering algorithm used by clicking on the "select clusters" button to explore alternative clustering outcomes. Clicking this button reveals a table containing clustering methods and their respective silhouette coefficients (a measure of the goodness of a clustering technique[151] ranging from -1 to 1). In this panel, we incorporate five clustering algorithms, including k-means clustering, hierarchical clustering, OPTICS clustering, DBSCAN clustering, and birch clustering[10, 23, 198]. PreFair by default displays the clustering method that produces the best silhouette coefficient in the Cluster View. We then apply either principle component analysis, multi-dimensional scaling, or t-Distributed Stochastic Neighbor Embedding (t-SNE) to project the clusters into two dimensions. By default, it displays the projection of the dimension reduction that comes with the highest preservation of variance. Users can click on the drop-down menu to switch the projection method. The generated clusters will dynamically update associated changes in the Radar View, Rule View and Tree View.

Radar View: The Radar View presents the centroid strategy of each cluster. Each radial spoke in the radar chart represents one of the 7 evaluation metrics with the range defined by the minimum and maximum values for each metric in the power set of pre-processing strategies. Each polygon represents a cluster. The radar chart facilitates at-a-glance comparisons of area, indicating general performance of the clusters according to the 7 metrics. This supports **DG 2** by enabling users to compare strategies' performance on all of the 7 metrics. The Radar View dynamically updates in response to changes in the Cluster View, Tree View, and Customization View.

Rule View: Although the centroid polygons in the Radar View provide a rough approximation of each cluster, we also want to provide a mechanism to characterize the extent of the most frequently occurring operations for the strategies in each cluster. Thus, we employ descriptive rules [147] to present the common characteristics held by the majority of the strategies in the same shared cluster. Each cluster's rule view is colored to match the encoding for the respective cluster in the Radar View. Clicking on the respective color swatches will generate the descriptive rule for the associated cluster. The Rule View explains, in the format of one descriptive rule, how the strategies in the selected cluster are different from the strategies outside the cluster. The rule is generated using Quinlan's C5.0 algorithm[147], a decision tree-based, optimized IF-THEN rule extractor.

For example, the Rule View in Figure 5.1D shows a rule containing three conditions. Each condition is presented in the form of a histogram (using the same color of the current cluster) which indicates the distribution of the rule for the inclusion of operations. These conditions describe the most frequently occurring operations that the majority of strategies in the currently selected cluster include, while most strategies outside of the cluster do not. Some clusters may have more than one rule condition, and all of the conditions can be collectively used to describe the given cluster's inclusion criteria. We rank these rule conditions by how well they differentiate the cluster in the measurement of information gain[106], and we show conditions in descending order in terms of added information gain in the Rule View.

5.4.3 Strategy Exploration and Comparison

Once users shrink the target search space, it is time to examine the performance of the remaining strategies and compare them to make their final decision on the appropriate strategy. We include the Tree View and Customization View to facilitate these goals, which primarily address **DG 2**.

Tree View: In the Tree View, we show the clustered strategies in the form of a word-tree structure, where common strategy prefixes are represented as a node in the tree that branch when strategies diverge. Clicking on the [regular] button will expand a collapsed node. Hovering on each node of the tree will dynamically highlight the current strategy's performance on the Radar View and the Parallel Coordinates View to show the predictive outcome if the strategy represented by the hierarchical sequence from root to hovered node were applied to the dataset. For instance, suppose the user hovers on the "PCA-5" node that has ancestor "L2 Normalization" under "Cluster 2". The Radar View would then show a polygon representing the model's performance if the pre-processing strategy (L2 Normalization $\rightarrow$ PCA-5) were applied to the dataset in pre-processing.

Customization View: Throughout their exploration of the search space of strategies, users may identify certain combinations of pre-processing operations they want to consider more deeply or directly compare to one another. They may be curious about a strategy that has been filtered out or want to compare the "what-if" scenario of a permutation of a preferred strategy. The Customization View supports the construction of pre-processing sequences whose model outcomes can then be visualized according to the resulting evaluation metrics in the Radar View. By clicking on the root node in the Customization View, users can choose the "Create Node" button to create an empty operation node, then drag and drop operations from above into the empty node. To continue building out a custom pre-processing sequence, users can continue building pre-processing sequences by iteratively adding leaf nodes. Users can delete a leaf node by selecting the "Delete Node" option. To visualize the custom strategy's performance, users can then click the "View Result" button, and the corresponding performance of the custom strategy will be shown in the Radar View.

	2	3	4	5	6
kmeans	0.86				
hac			0.39		
dbscan					0.39
optics		0.52			
birch				0.39	

Figure 5.5: Usage scenario 1 (subsection 5.5.1). **a** Clipping the range of statistical parity difference and error rate ratio. **b** The centroid of cluster 2 has decent performance other than error rate ratio. **c** Switch to another clustering method in the Cluster View. **d** Observe rules among the three clusters and infer disparate impact remover algorithm may positively impact the error rate ratio for this prediction task.

5.5 Usage Scenarios

In this section, we introduce two usage scenarios to demonstrate how PreFair can assist users in making fairness-informed decisions for data pre-processing.

5.5.1 Searching with Prioritized Metrics

Consider Darren, a data science engineer, is using PreFair to analyze loan applications in the German Credit dataset[58] to determine creditworthiness. He utilizes PreFair to facilitate his decisions and nudge his behavior in order to make responsible choices for pre-processing his data. The dataset consists of both categorical and numerical data and contains 1000 instances (rows) described by 20 attributes (columns). After deciding to use an Adaboost classifier[155] with 100 estimators and split the train/test set with a 7:3 ratio to predict loan approval, he wants to deter-

mine what data pre-processing strategy he should use so it honors model fairness. Darren prefers to prioritize "statistical parity difference" and "error rate ratio" while balancing choices that will result in decent values for other metrics.

He first looks into the Parallel Coordinates View to scan through the overall distribution of potential options. He leaves all of the operation buttons selected to begin his search with the full set of possibilities. Darren creates a filter on the Parallel Coordinates View by dragging along the statistical parity difference axis, and observes that if he sets the filter in the range [-0.1, 0], about half of the options are excluded. If he further shrank the range to [-0.05, 0], only a few strategy options would remain. While exploring, Darren makes a similar observation when filtering the error rate ratio into a relatively small range. Darren ultimately chooses to narrow down the search space by clipping the range of statistical parity difference to [-0.13, 0], and the range of error rate ratio to [0.08, 0.15] (**Figure 5.5a**). Given the numerous strategies, Darren wants to see how clustering can help further shrink the search space. He moves to the Cluster View, where he can visualize the filtered strategies in a 2-dimensional scatter plot. Each point in this plot represents a pre-processing strategy, with colors indicating the cluster to which the strategy belongs. The Cluster View's scatter plot nudges Darren's behavior by providing a visual summary of how strategies group based on their overall performance. By identifying dense clusters, he can infer that the strategies within these clusters are likely to perform similarly and focus his attention on specific groups rather than individual strategies.

After clicking on the "Run Clusterings" button on the Parallel Coordinates View, a scatter plot with colors representing three resulting clusters is shown in the Cluster View. Looking at the scatter plot, Darren sees each strategy represented by a point, colored by cluster. Darren can see one cluster has numerous possible strategies that are similar (tightly grouped in the plot), while another has few.

Darren then examines the radar chart to identify which cluster has the proper-

ties he cares about. The radar chart presents the centroid strategy of each cluster, showing performance across all metrics. Darren notes that although cluster 2 has great performance on statistical parity difference and the majority of other metrics, its error rate ratio performance is the lowest among the three clusters (**Figure 5.5b**). Given he cannot find a cluster with ideal scores for his priority metrics, he realizes the default clustering method may not necessarily help him to narrow down the search space.

To ensure his behaviors and decisions are responsible with optimal fairness outcomes, Darren decides to swap the clustering method. Clicking on the "select clusters" button, a clustering method selection table pops up, and Darren observes the silhouette coefficient of each clustering option is provided. He switches the method from the default, k-means clustering with 2 clusters, to OPTICS clustering with 3 clusters (**Figure 5.5c**), which has the second highest silhouette coefficient value. The Cluster View, Radar View, Rule View, and Tree View are accordingly updated.

In the Radar View, Darren observes that the centroid of cluster 3 has the best error rate ratio compared to other centroids, and it also has a decent statistical parity difference. This looks promising. The only problem with cluster 3 is the centroid strategy has a disparate impact that is below average. Darren realizes the error rate ratio of the other two clusters is dramatically worse than cluster 3, and wonders why that would be the case, and in general what operations can bring a better error rate ratio for this dataset. He looks into the Rule View, which complements this by providing descriptive rules that explain the common characteristics of strategies within each cluster, enabling Darren to understand why certain strategies perform better in specific metrics. Darren clicks on cluster 3's button in the Rule View. A rule list shows that the major characteristic of strategies in this cluster is that most of them include the re-weighing algorithm. Darren observes the rule lists of both cluster 1 and cluster 2 do not contain the disparate impact remover algorithm, while cluster

3 does (**Figure 5.5d**). He thus infers that applying the disparate impact remover algorithm may positively impact the error rate ratio for this prediction task.

Based on the characteristics of cluster 3, Darren wants to get rid of strategies whose disparate impact is below average. Therefore, Darren adjusts the filters in the Parallel Coordinates View to set the statistical parity difference to be above -0.075, the error rate ratio to be above 0.125, and the disparate impact to be above average. After that, Darren clicks the "Run Clustering" button again to output a refined search space.

Now Darren observes there remain 13 qualified strategies. Darren inspects the Radar View and finds that although the centroid strategies of the three clusters are now different, their difference is not as dramatic as in the first round of clustering. After carefully comparing the clusters, he reflects there is no clearly superior cluster and decides to examine all 13 remaining strategies in the Tree View. He clicks the [regular] button and observes the strategies in each word-tree cluster. Darren hovers through each leaf node in turn to examine the performance of the pre-processing strategy on the Radar View. After thorough consideration, Darren believes that choosing L1 Normalization → Disparate Impact Remover is his final choice for pre-processing his data, and he is confident that such a decision is responsible after careful iterations with this DS intervention tool.

5.5.2 Strategy Brainstorming

Abbey, an undergrad student in a data science class, is using PreFair as an intervention tool to understand and navigate her behaviors during the process of learning how different pre-processing strategies can affect model outcomes and how some algorithms can reduce model bias at the pre-processing stage. She first wants to see how normalizing data in different ways can influence model performance. Starting with the Customization View, she clicks on the root node then drags and drops "L1

Normalization" into the empty node. Clicking on the "View Result" button, the performance of this pre-processing strategy is then shown in the Radar View. After seeing the performance with L1 Normalization, she drags and drops "L2 Normalization" on the leaf node to replace the L1 Normalization. She presses the "View Result" button again. Compared to L1 Normalization, she observes that disparate impact increases a bit and precision decreases a bit, while other metrics do not change dramatically. Abbey next wants to explore how bias mitigation algorithms can influence model performance. She adds another node after "L2 Normalization", and drops "Disparate Impact Remover" onto the leaf node, followed in turn by replacing it with "Re-weighing Algorithm." Abbey notices that each of these two bias mitigation pre-processing algorithms can improve some fairness metrics: applying Re-weighing algorithm can mitigate statistical parity difference by 0.12 and Disparate Impact Remover can dramatically reduce disparate impact by 0.17. In the context of the German Credit Dataset, Abbey observes that Disparate Impact Remover slightly out-performs Re-weighing algorithm for enhancing model's overall fairness performance, given that the data is first processed with L2 Normalization.

5.6 Preliminary User Feedback

In order to understand how the PreFair approach helps users make fairness-aware data pre-processing decisions, we collected feedback from six machine learning experts. Each session lasted 45-60 minutes and consisted of a brief introduction to the project, a background questionnaire, a scripted walk-through of the prototype, task operation, and a post-study survey. Participants volunteered for the study and were not compensated. The primary goal was to understand how well PreFair helps the user (1) narrow down the search space of pre-processing strategies, and (2) assess fairness trade-offs of different pre-processing strategies. In this section, we provide

the participant's background, describe the methodology, and report our findings.

5.6.1 Participants

We recruited six participants who had 4+ years of experience designing and applying machine learning solutions. Participants were recruited via email and through the authors' personal Twitter network. Table 5.1 summarizes participants' information. The participant pool had a diverse background for categories such as – *industry* (3 academic, 3 industry), *gender* (4 male, 2 female), *age* (range: 27-35; median: 29), and *degree* (Ph.D.: 2, MS: 1, Ph.D. students: 3). The expertise of the participants spanned over multiple machine learning sub-domains such as – *Machine Learning Fairness* (P1), *Natural Language Processing* (P2, P4), *Data Science* (P3, P6), and *Applied Machine Learning* (P5).

All of our participants have a background in a computational field (e.g., Computer Science, Data Science, Human-Computer Interaction). P1 is a Ph.D. student who works in the area of machine learning fairness; P2 and P4 are Ph.D. students working on Natural Language Processing; P3 and P6 are data scientists, and P5 is an applied scientist working on a machine learning related area. None of the participants had previously used any visual analytic tool to help mitigate bias in machine learning.

	P	Y	E	D	G	A
P1	Academia	7	Y	Ph.D. (IP)	F	28
P2	Academia	4	N	Ph.D. (IP)	M	27
P3	Industry	5	N	Ph.D.	M	31
P4	Academia	5	N	Ph.D. (IP)	F	28
P5	Industry	7	N	Ph.D.	M	35
P6	Industry	8	N	M.S.	M	30

Table 5.1: Participants' information about **P** (Profession), **Y** (Years of experience in machine learning), **E** (machine learning fairness Expert or not), **D** (highest degree attained or in progress (IP)), **G** (Gender), **A** (Age).

5.6.2 Method

We performed our study as a hybrid in-lab (5 participants) and virtual (1 participant) study which utilized the Zoom software allowing for remote collaboration such that the application ran locally on the administrator's system and the participants interacted remotely with it. This remote access was achieved by screen sharing and remote control.

The study commenced with a background questionnaire designed to gather information on various aspects of participants' experiences and perspectives. This questionnaire collected details about their domain experience, familiarity with algorithmic fairness, the types of data they typically work with, their usual data pre-processing routines, and their views on the importance of the data pre-processing stage in influencing the fairness of machine learning models. We gave each participant a tour of the tool by introducing each component of the prototype and interactions between each panel. Next, we asked participants to practice and interact with the prototype, followed by an opportunity to ask questions.

After each participant became familiar with the prototype, they completed the primary task of the study: *"Given the German Credit Dataset, suppose you decided to use an Adaboost classifier to predict applicants' loan approval. You want to ensure the model's predictive fairness on young female applicants while maintaining a decent predictive performance. Use* PreFair *to select an optimal pre-processing strategy, in your mind, given that you want to prioritize statistical parity difference and disparate impact."*

We encouraged participants to think aloud and ask questions if they got stuck during the session. After each participant finished the task, we concluded with a feedback session (in the format of a semi-structured interview) where we asked them about their experience using PreFair. Lastly, participants were requested to furnish retrospective feedback concerning the usability of the prototype and its features via

a post-study survey. The post-study survey focused on the system's overall usability, the utility of each specific view, and their combined effectiveness in supporting the completion of the assigned study task. A critical aspect we explored was how a system like PreFair integrates into the participants' data science project pipelines. Furthermore, the post-study survey aimed to understand their perspectives on the role of pre-processing in enhancing fairness in machine learning applications, specifically asking for their insights on the importance of the data pre-processing stage in influencing model fairness after using the PreFair system. To complement this, we also asked whether using the PREFAIR system had enriched their understanding of the significance of the data pre-processing stage in ensuring fairness in machine learning. Sessions were recorded to analyze the interactions and user experience.

5.6.3 Qualitative Findings

Based on our observations during the study task and the post-study interview, we summarize key insights below.

PreFair helped people identify trade-offs and explore options. Overall, participants indicated satisfaction with PreFair. P6 said PreFair provided *"utility to find good strategies among a large number of candidates."* Similarly, P1 commented on the cascaded workflow: *"Parallel Coordinates View helps you do the first round of filtering. Cluster View does a second round (of filtering), and the other views help with the last round of searching."* We also noticed that the Parallel Coordinates view was considered the most useful view by all participants (verified by usability scores, included in supplemental materials): P2 stated that *"overall, the most useful information is all in the PC view (...) it contains almost 90% of the information I need for shrinking the option space."* Beyond that, participants found the Parallel Coordinates View helped them to understand the performance distribution of evaluation

metrics. P6 commented it is *"more than a filter (...) By sliding the filters, I can understand for a range of one metric, how many strategies have good performance in other metrics I'd have to sacrifice for."* P5 similarly emphasized that the PC view shows *"the density of lines (which) tells me that the strategies' performance is unevenly-distributed, and I will bear it in mind when I move the filters."* P4 found that adjusting the filters in the Parallel Coordinates View helped narrow down the options without losing too many candidate strategies: *"According to the PC view, I'm able to apply filters only on the important metrics that I should pay attention to. At the same time, I can decide the selection range of these two filters according to the changing trend of other metrics (...) I can get a sense of how much I would set the filters (without losing too many candidates)."*

The value of the interface is highly dependent on the quality of the clustering outcome. 5 out of 6 participants commented on the cascaded workflow to narrow down the search space of pre-processing strategies. Accordingly, many participants emphasized that the quality of filtering in the Parallel Coordinate View is essential for the success of the pipeline. P4 indicated: *"the clustering quality matters a lot in order to deliver a good second round of filtering. I get the idea that you realized this as well. You made several clustering options in case one single clustering does not work."* This was echoed by P6 as well: *"If I can not find a clustering method that looks good enough to separate clusters, I would further change the PC filters and re-do the clustering."*

Some views were complementary but had overlapping functionalities. During the study, we observed participants were satisfied with the views we designed to support each other, as P1 commented: *"PC and all other views are complementary."* P3 stated: *"For the last round of searching, I basically clicked on each dot in the Cluster View and saw their performance in the Radar View. I like the design of the*

radar as I can easily see the difference by comparing the area size, but for sure I kept an eye on the prioritized metrics as well." However, we also observed many unintended workflows that demonstrated the overlapping functionalities of some of the views, given that participants had varied approaches in making the final decision. For instance, before the participants made their final decision, P3 and P5 mainly relied on viewing the strategy's performance from the Parallel Coordinates View by clicking on the points in the Clustering View, whereas P1, P2, P4, and P6 primarily relied on the Radar View to make the same comparison. Some preferred one view over the other for such comparisons due to inherent limitations. For instance, P1 shared *"I can imagine when there are a lot of strategies, the PC view will be less efficient for viewing the selected strategy while the radar plot can do a better job of clearly viewing the selected strategy's result."*

Rule View was seldom incorporated into the workflow. Only P1 used the Rule View while completing the given task. P1 expressed that the Rule View had the potential to add interpretability to the favored cluster chosen by users, stating: *"the cluster and Rule View provides some descriptive info about why some strategies perform better."* P2 did not use the Rule View during the study, but could envision its utility in another scenario: *"If the goal is to summarize a profile of successful strategies, then the Rule View will be super helpful."*

Customization view was seldom used, but users could imagine useful scenarios. Apart from the Rule View, we also found participants seldom used the Customization View as we had expected. Only one out of six participants used the Customization View during the study. P2 indicated that this view is *"not directly useful"* for completing this task; P4 mentioned that they *"did not think of using Customization View during the study at all."* Nonetheless, P3 suggested that the Customization View would be helpful in certain scenarios: *"I can imagine it could*

be useful for those who already have some plan in mind to deal with data. They can grab and use it to test their ideas. It has high flexibility for customized pipelines of pre-processing strategies."

Users wanted a better way to maintain awareness of their provenance. In spite of multiple ways to view the performance of pre-processing strategies, participants noted several opportunities for the interface to do a better job helping them keep track of their process. P1 suggested, *"It would be great if you have a bookmark function, so I can save some good candidates (strategies) for future comparison."* P6 wanted to use a bookmark function but ended up finding an alternative way given the Tree View: *"I also used the hover function as a kind of bookmark. When I see a good candidate, I hover on it in the tree view and keep clicking on other points in the Cluster View, so I can compare the strategy that is bookmarked (hovered in the tree view) with the one I select."* P2 indicated a desire for the interface to track the number or percentage of strategies that are kept after filtering: *"I want to have an information board in the Parallel Coordinates View, telling me given my current filtering, how many strategies are there in total, how many are selected, and how many are filtered."* P3 suggested another improvement to track their process in the Parallel Coordinates View: *"Adding an average point on each axis in the PC View would be great for benchmarking how many strategies I got rid of while filtering."*

Users envisioned scalability solutions for practical use. Many participants shared their opinions on how PreFair could be adapted into real-world machine learning model deployment, namely with respect to dealing with larger dataset applications. P2 suggested scalability challenges could be addressed by an ensemble approach: *"you could sample small batches from the whole dataset and use this to select pre-processing strategies for each small batch. With all strategies picked from small batches, you could make decisions on how to process data on the large dataset*

in an ensemble way (like using strategies you chose in small batches to do majority voting for the bigger dataset).” P6 suggested that the idea of filtering and narrowing down the search space could also be applied for model parameter grid search: *“If we could use* PreFair *for comparing all parameters we want to try, you could calculate their (different parameter combinations’) performance for all metrics that we care about, and make decisions with a similar mechanism.”* Apart from the comments on scalability, we also receive comments on integrating PreFair into frequently used machine learning deployment environments. P3 suggested *“It would be interesting to see if you can deploy this system into Jupyter Notebook or Visual Studio Code as a Python package or add-on.”*

5.6.4 System Improvements

In response to user feedback during the summative evaluation, we engaged in one additional round of iteration on system development. As pointed out by P1 in the qualitative user feedback, the Parallel Coordinates View may not scale when there are a large number of strategies. To mitigate this, we implemented a default axis ordering on the Parallel Coordinates View using the algorithm introduced by Heinrich et al. [87] to minimize edge crossings and improve readability of the chart. The system code, code for the experiment in section 5.2, and materials related to user study (survey, questionnaire, scripts, and results) are included in supplemental materials[1].

5.7 Discussion

We anticipate that PreFair may be useful both in the context of machine learning engineers building models for applications in government or industry as well as an educational tool to teach students the potential fairness impacts of seemingly benign

[1] https://drive.google.com/drive/folders/13lcHLH6n-nitiJI3ufjqIbNtD3o8Gp7L?usp=sharing

data pre-processing sequences. Below we discuss limitations of our current approach as well as avenues for future work.

Scalability and Generalizability of the PreFair Technique. While PreFair currently concentrates on the pre-processing stage of the machine learning pipeline as a proof-of-concept, we acknowledge that bias may manifest and intensify at various stages throughout the pipeline even if the data appear to be "good" at the pre-processing stage. We can thus envision a more robust application of the PreFair technique that incorporates in-processing and post-processing stages to represent a holistic, end-to-end tool for assessing model fairness through the permutation of a multiverse of model choices[81]. This could include data wrangling, model selection, training, and a thorough assessment of model outcomes. The technique can also be extended in a straightforward fashion to support alternative evaluation metrics, model types, and datasets (including multimedia training data such as text, images, etc). For example, the system could encompass a mix of model types and training epochs, allowing PreFair to refine the search space for solutions that span the entire machine learning pipeline. Allowing a selection of model types would necessitate incorporating options for model hyper-parameters as well. This would present challenges in PreFair due to the numeric nature of typical parameter search spaces, as opposed to categorical options for including pre-processing steps. However, autoML techniques could be used for hyper-parameter tuning via grid search behind the scenes, prior to feeding the performance data into PreFair. Using autoML to optimize for accuracy would not undermine the concept of fairness from choosing pre-processing strategies, because there is already an assumption that the model performs as accurately as possible, but that we still gain fairness benefits from the strategy selection. Certainly, autoML could be applied to other metrics to tune as well, e.g., targeting fairness explicitly or in an ensemble.

How much does it matter to compare pre-processing strategies? Prior

work[24], validated by our analysis (Figure 5.2), suggests that changes in pre-processing operations and even ordered sequences of operations can affect model fairness. On the other hand, our preliminary feedback came from 6 expert participants, none of whom indicated that they explore alternative pre-processing strategies or consider its effects on model fairness. This led us to revisit our assumptions about data science practices. There are numerous reasons that could explain this trend, e.g., some experts may be working on applications wherein fairness is not viewed as critical, or, experts who do consider model fairness may do so at different stages of the machine learning pipeline (e.g., in post-processing[5, 114]). Additional work is needed to investigate the efficacy of addressing model fairness from the pre-processing stage vs. considering downstream mitigation strategies. In cases where model builders are unaware or ambivalent about model fairness, we can investigate future interventions to increase the investment in fair model outcomes.

Human-in-the-Loop Fairness. In computational fields, there is naturally impulse to operationalize, quantify, and compute. In many scenarios with well-defined problems, these approaches are ideal to achieve efficient solutions. However, applications that are less crisply defined or those of ethical importance may ultimately still benefit from human-in-the-loop oversight. Auditing models for fairness is one such applications where social justice and other matters of societal importance are at stake; hence, maintaining human-in-the-loop oversight is critical. While PreFair provides an approach to explore fairness trade-offs, it still solely relies on quantified metrics to support this oversight. Future work may explore approaches that nudge or incentivize model builders to directly engage with communities affected by models or consider integration of prototype components that will allow more flexible exploration of model fairness, e.g., via supporting what-if analysis.

Optimizing Pre-processing Strategy Selection. The current interface for PreFair emphasizes user agency to explore the space of pre-processing strategies and

select the option that best fits their priorities. If prioritization of the evaluation metrics is known, an automated optimization strategy can solve the problem of choosing the appropriate pre-processing strategy. In situations where users may not have specific criteria for evaluation until they explore the space, an interactive visual analytic approach such as PreFair is suitable. However, user feedback from the evaluation revealed opportunities to further reduce the burden on the user to evaluate alternative strategies. If we assume that (1) the user has filtered out strategies that do not contain the appropriate pre-processing operations for their application and (2) that all of our evaluation metrics have defined "preferred" values (e.g., higher accuracy is better; lower fairness metric value is better), then we can reduce the space of remaining options by considering the pareto front[115]. That is, users need not consider the full range of all solutions, because some may be strictly inferior to others according to the evaluation metrics. For instance, the pre-processing strategy A $\rightarrow$ B may be dominated across all evaluation metrics (strictly inferior in fairness and performance metrics) compared to B $\rightarrow$ A. In this case, dominated strategies like B $\rightarrow$ A could be pruned from the analysis. P1 made a similar suggestion: *"If I can set a benchmark strategy, it would be great if I click on a button, all strategies that are better than the benchmark are kept, and others will be grey. This will make the process smoother."*

5.8 Limitations

The development and implementation of PreFair revealed several limitations that inform our broader thesis on behavior change interventions for responsible data science. While PreFair successfully demonstrates how technical tools can support responsible pre-processing practices, it primarily addresses the "capability" aspect of our behavior change framework without fully engaging with motivation and opportunity factors identified in chapter 3. The tool assumes users are already motivated to

pursue fairness goals but may lack the technical capabilities to evaluate pre-processing strategy impacts. Additionally, our evaluation focused on technical effectiveness rather than behavioral outcomes, making it difficult to assess PreFair's impact on actual data scientists' practices over time. This limitation highlights the critical need for comprehensive evaluation frameworks that measure not just technical capabilities but behavioral changes in real-world contexts. Future work could integrate PreFair-like tools into broader intervention systems that address motivational factors and workflow compatibility, while employing structured evaluation frameworks to assess intervention effectiveness. These insights directly inform our subsequent research on evaluating intervention efficacy

Within the scope of **RQ3**, PreFair has focused on just one phase of the data science workflow, suggesting that additional interventions may be needed to effectively address responsible practices across other stages of the data science lifecycle. It focuses on the pre-processing stage of data science workflows—while important—captures only a portion of the complex decision-making processes that impact responsible data science outcomes. The tool's design does not allow us to quantify intervention efficacy and success thoroughly across the entire data science pipeline, from problem formulation through deployment and monitoring. This limitation highlights the need for more comprehensive evaluation methods that can measure intervention efficacy across the entire data science lifecycle, prioritizing approaches with quantifiable metrics and expanding beyond fairness metrics to include a broader set of evaluation methods.

5.9 Summary

In this project, we designed and implemented PreFair, an interactive intervention tool to provide insights into **RQ 3** (*How effective are behavior change interven-*

tions at improving responsible data science outcomes?). We first presented a machine learning experiment to showcase that model pre-processing influences DS model fairness in a notable way. Second, we elaborated on the PreFair prototype by introducing its design process and goals, interfaces, and usage scenarios. Finally, we described the user study and its preliminary evaluation results with DS experts, suggesting that PreFair is a promising approach to help the DS model builders understand fairness trade-offs when pre-processing data. The results show that successful interventions should be designed with specific, measurable metrics in mind—an area where PreFair excelled through its visual representations of fairness metrics. However, our findings also reveal that effectiveness depends on how well interventions integrate with data scientists' existing workflows—a limitation we encountered with PreFair's implementation. This tension highlights a critical gap in the field: the need for intervention efficacy measurements from both outcome and process perspectives. This realization motivates our next chapter, where we evaluate intervention efficacy that provides the missing structure by establishing concrete performance goals for intervention designers.

Chapter 6

Evaluating Behavior Change Interventions for Responsible Data Science

6.1 Motivation

We now want to address a critical gap in responsible data science research: the evaluation of behavior change intervention efficacy. Despite growing interest in developing interventions to promote responsible practices, systematic approaches to measuring their effectiveness remain underdeveloped. Current approaches to evaluation in RDS often focus either on technical metrics (such as fairness indices) or qualitative user experiences alone, but rarely integrate these perspectives into comprehensive evaluation paradigms. This fragmentation makes it difficult to systematically improve interventions or to compare their efficacy across different contexts and user populations. Moreover, the field lacks consensus on what constitutes "successful" intervention adoption.

After establishing a framework for RDS through the lens of behavior change theo-

ries in chapter 3 (addressing **RQ1**), proposing a design space for behavior change interventions for RDS in chapter 4 (addressing **RQ2**), and developing a behavior change intervention tool designed for RDS in chapter 5 (addressing developing an intervention in **RQ3**), we seek to answer the second part of **RQ3** (evaluating intervention efficacy) and bridge the gap between theoretical aspirations and practical challenges in fostering responsible behavior in data science workflows. Specifically, we focus on evaluating the efficacy of behavior change interventions (BCIs) designed to encourage practices such as fairness-aware modeling, bias mitigation, and comprehensive data exploration. In this project, **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.** Rather than testing PreFair as an intervention in this study, we opted for the two alternative interventions: (i) a simpler priming approach and (ii) an established fairness toolkit (Aequitas [97]). A key reason for this choice is that PreFair is an ex-situ intervention that requires extra effort to incorporate into the data science development environment. We suspect that in-situ interventions (embedded directly within the notebook environment itself) would likely be more effective and promote responsible behaviors and outcomes, making them a more appropriate first intervention to test. 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. This work[56] is currently in preparation for submission to CHI 2026.

6.2 Methods

This user study is designed to assess the efficacy of behavior change interventions in promoting responsible data science practices. By comparing different levels of intervention, this study aims to explore how data scientists' approaches to fairness, data exploration, and bias mitigation are influenced under varying conditions. This section presents an overview of the study objectives, the tasks that participants were asked to perform, and the chosen interventions.

6.2.1 Tasks & Interventions

The study design (depicted in Figure 6.1) 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. 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 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 task in a basic Google Colab notebook without instructions on responsible data science or model fairness.
2. **Prime**: Instructions are augmented to prime participants to be concerned with model fairness. For example, we added the following text 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. Within the context of the design space from Chapter 4, **Aequitas** [97]: 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. Aequitas exemplifies an intervention that enhances both Capability and Motivation within the COM-B model (What) for experienced data scientists (Who), functioning 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 first completed one of the two tasks in the **Control** condition, followed by the second task using one of the interventions (**Prime** or **Aequitas**, randomly assigned). After finishing each task, participants were asked to complete the NASA-TLX[86] questionnaire to assess their cognitive load during the process. Once participants finished the **Prime** or **Aequitas** phase, researchers collected their qualitative feedback on how the intervention influenced their decisions, attitudes, and behaviors compared to the **Control** condition and their responses to our interview questions.

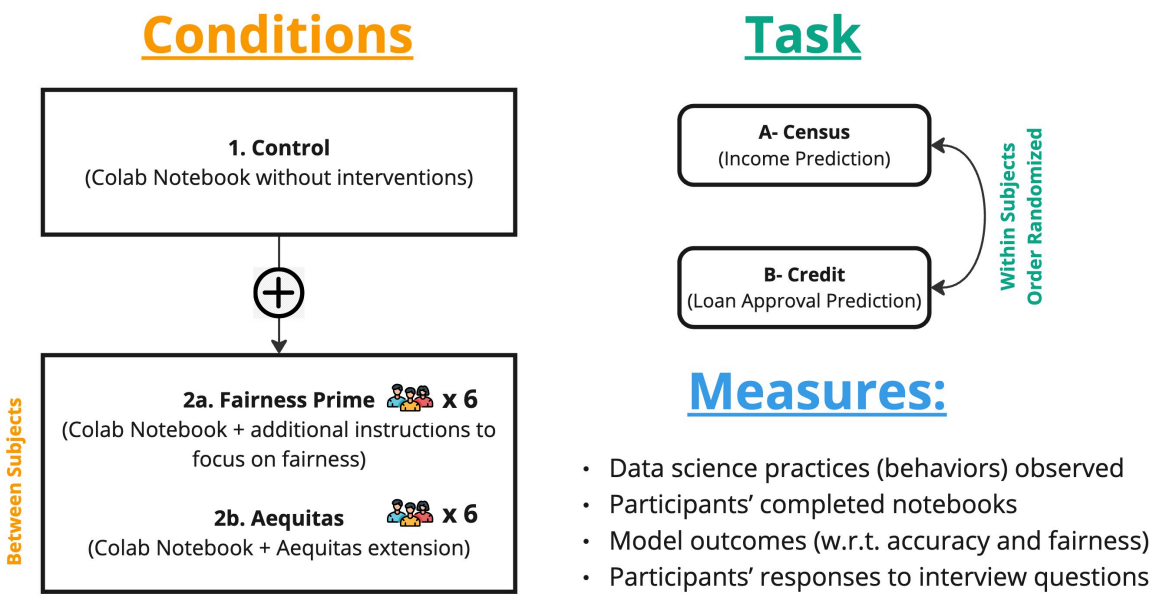

Figure 6.1: 6 participants were recruited to conduct Task 1 and 2a, with the other 6 participants tasked with Task 1 and 2b. with each task being randomized and presented in the Census dataset or Credit dataset context.

6.2.2 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 invitation on Linkedin to data science practioners in the industry. 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 6.1.

6.2.3 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 effect from earlier

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

Table 6.1: The participants' demographic information on their gender, year of experience(YOE), age, and Occupation

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. All participants provided verbal responses to our interview questions, with the exception of one participant (P7) who shared his responses via email due to recovery from recent throat surgery. Participants were provided with a Colab Notebook containing dataset characteristics and task instructions corresponding to their condition in the study. All AI features including Gemini were disabled.

6.2.4 Responsible Data Science Practices and Data Collection

To evaluate the efficacy of behavior change interventions, we establish a 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.[46], which outlines the data science process, and Bellamy et al.[19], 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[62], 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[48], 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[144].

6.2.5 Hypotheses

In this user study, we formulated several hypotheses regarding the impact of behavior change interventions on responsible data science practices:

H1: **Responsible Behaviors** - Both Prime and Aequitas interventions will promote responsible behaviors compared to the Control, and Aequitas will outperform Prime.
H2: **COM-B Factors[128]** - Both Prime and Aequitas interventions will primarily influence the 'Motivation' factor within the COM-B framework, compared to 'Capability' and 'Opportunity'.
H3: **Model Fairness** - Both Prime and Aequitas interventions will improve fairness metrics of the resulting models, and Aequitas will outperform Prime.
H4: **Model Performance** - Both Prime and Aequitas interventions will not significantly affect the accuracy of the resulting models.
H5: **Cognitive Load** - Both Prime and Aequitas interventions will not significantly increase the cognitive load compared to the Control.

6.2.6 Measures

To evaluate the efficacy of behavior change interventions, we employ a mixed-methods approach combining quantitative analysis and qualitative insights. 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). These quantitative measures are also supplemented with qualitative feedback from participant interviews, including reflective quotes on their decision-making processes, challenges, and perceived answers to our interview questions. We demonstrate the statistical measurement for each hypothesis below:

1. *H1:* We conducted 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.
2. *H2:* We utilized 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.
3. *H3:* A repeated-measures t-test was performed to compare fairness metrics (false discovery rate ratio of disadvantaged and non-disadvantaged groups) across conditions (Control, Prime, and Aequitas).
4. *H4:* A repeated-measures t-test was performed to compare model's accuracy across conditions (Control, Prime, and Aequitas).
5. *H5:* We applied 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.

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 6.2: 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%**	**83%**	**100%**	**67%**	**17%**	**33%**	**100%**	**83%**	**100%**	**100%**	**100%**	**0%**	**8.2**

Figure 6.3: An overview of the participants' responsible behaviors within Aequitas group's control and treatment sessions.

6.3 Results

6.3.1 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 6.2.4 (maximum of 12 responsible behaviors).

The data reveals a clear pattern of a larger number of responsible behaviors observed while using the interventions compared to the control, we report the overviews of observed responsible behaviors within the Prime and Aequitas groups within Figure 6.2 (Prime) and Figure 6.3 (Aequitas). When exposed to the Prime intervention, participants exhibited an average of $\mu = 5.7$ ($\sigma = 0.75$) responsible behaviors compared to their Control behaviors of $\mu = 2.8$ ($\sigma = 0.69$). Using a repeated measures t-test, we determined this to be statistically significant ($t = -6.26$, $p < 0.01$). Those

using the Aequitas toolkit demonstrated the highest average at $\mu = 8.1$ ($\sigma = 1.34$) responsible behaviors compared to their Control behaviors of $\mu = 3.67$ ($\sigma = 1.25$), which we also found to be statistically significant ($t = -9.22$, $p < 0.01$).

It is worth noting that we examined the baseline responsible behaviors of participants in the control condition across the Prime and Aequitas groups to ensure initial comparability. A statistical comparison yielded a result of ($t = -9.21$, $p = 0.09$), suggesting 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]."* In contrast, participants using Aequitas described more concrete shifts in practice. 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."* 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."*

6.3.2 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 6.2). 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 Positive Influence, 2 = Positive Influence) from participants on their perceived Capability, Opportunity, and Motivation across the three conditions (Table 6.2). 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 ($W = 21$, $p = 0.01563$), indicating that Motivation ratings were significantly higher; however, the comparison of Motivation to Opportunity was not significant ($W = 3$, $p = 0.07865$). 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 ($W = 3$, $p = 0.07865$) and Opportunity ($W = 3$, $p = 0.07865$), the results were not significant. 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

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?

Table 6.2: Interview questions we asked participants after they finished the treatment study session.

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 still played a role, participants largely perceived Motivation as the driver of behavior change.

6.3.3 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 ($\mu = 2.05$, $\sigma = 0.39$), with a paired t-test showing a non-significant improvement ($t = 1.42$, $p = 0.1075$). In contrast, participants in the Aequitas condition demonstrated a substantial improvement in fairness ($\mu = 1.18$, $\sigma = 0.16$) over their Control fairness ($\mu = 2.45$, $\sigma = 0.31$), which we found to be statistically significant ($t = 9.2$, $p < 0.01$). 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**: Both Prime and Aequitas improve fairness (though Prime was not a statistically significant improvement), and 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."*

6.3.4 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. The Prime group (n=6) had an average accuracy of $\mu = 0.60$ ($\sigma = 0.03$) compared to their Control accuracy of $\mu = 0.59$ ($\sigma = 0.05$), which was not significantly different ($t = -0.59$, $p = 0.58$). Similarly, the Aequitas group (n=6) had an average accuracy of $\mu = 0.58$ ($\sigma = 0.04$) compared to their Control performance of $\mu = 0.61$

($\sigma = 0.02$), which was not significantly different ($t = 1.59$, $p = 0.17$). We thus find **support for H4**: that neither the Prime nor Aequitas interventions significantly affect model accuracy.

These findings have important implications for the responsible data science field. The fact that neither intervention significantly affected model accuracy suggests that making models more responsible need not be a "zero-sum game." In other words, practitioners do not necessarily need to sacrifice technical performance to achieve greater fairness. This challenges the common perception that there *must* be a trade-off between ethical considerations and model performance. While our study focused on specific datasets and interventions, these results provide promising evidence that with appropriate tools and motivation, data scientists can improve fairness outcomes while maintaining model quality.

6.3.5 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 6.3 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 ($\mu_C = 0.83$, $\mu_P = 1.17$, $W = 1$, $p = 0.16$). 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 ($\mu_C = 0.67$, $\mu_A = 0.67$, $W = 3$, $p = 0.5$). 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** and that Aequitas *does* increase cognitive load.

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

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

Table 6.3: 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*.

Hypothesis	Description	Statistical Result	Conclusion
H1: Responsible Behaviors	Both Prime and Aequitas interventions will promote responsible behaviors compared to Control, and Aequitas will outperform Prime.	Prime vs. Control: p < 0.01 Aequitas vs. Control: p < 0.01	Supported
H2: COM-B Factors	Both interventions will primarily influence the 'Motivation' factor within the COM-B framework.	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 significantly affect model accuracy.	Prime vs. Control: p = 0.58 Aequitas vs. Control: p = 0.17	Supported
H5: Cognitive Load	Neither intervention will significantly 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 6.4: An overview of the outcomes for the 5 hypotheses we proposed

task.

6.3.6 Summary of Results

We summarize the outcomes for the hypotheses in Figure 6.4. 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.

6.4 Discussion

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.

On the Role of Motivation: 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. P9 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.

Balancing Cognitive Load: 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, P1 remarked, *"It [Aequitas] is like a forcing function to let me revisit my model development to check and refine my model deployment."* This trade-off between efficacy and usability suggests that future fairness tools must prioritize intuitive design and workflow integration to reduce cognitive burden. Techniques such as interactive visualizations or automated fairness suggestions could mitigate these challenges while retaining the benefits of technical tooling.

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.

6.5 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[92].

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) [140]. 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.

6.6 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 having an 0 as a neutral op-

tion) 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. 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 examine whether the observed behavior changes persist over time and how they evolve as practitioners gain familiarity with interventions like Aequitas.

6.7 Summary

In this project, we addressed **RQ3** (*How effective are behavior change interventions at improving responsible data science outcomes?*) by evaluating the efficacy of behavior change interventions for responsible data science through a user study with 12 data scientists. Our findings demonstrate that both interventions (Prime and Aequitas) increased responsible behaviors, with Aequitas significantly improving fairness metrics while maintaining model accuracy, though at the cost of higher cognitive load. Prime effectively boosted motivation without additional cognitive burden but showed limited impact on fairness outcomes. These results reveal a critical distinction between improving behaviors and improving outcomes in responsible data science—while behavior changes are necessary precursors, they don't guarantee improved fairness metrics, suggesting future interventions should bridge this gap by creating sustainable behavior changes that reliably produce responsible outcomes.

Chapter 7

Discussion

This dissertation has explored the integration and application of behavior change theories into responsible data science practices, addressing a critical gap in current approaches to AI ethics and fairness. Based on this novel perspective, I share several discussion points for advancing responsible data science by targeting the human factors that drive the ethical decision-making and behaviors in AI development and implementation.

7.1 The Complementary Nature of Technical and Behavioral Approaches

This dissertation introduces a novel perspective on responsible data science by highlighting the critical but often overlooked role of human behavior. While previous research has predominantly focused on technical solutions—such as algorithm modifications and bias mitigation strategies—our work demonstrates that these approaches alone are insufficient. By adapting established behavior change frameworks from cognitive and clinical psychology, I have shown that responsible data science requires addressing both technical systems and human behaviors. This complementary

approach recognizes that even the most sophisticated algorithmic solutions remain susceptible to human biases during their development and implementation. This integration of technical and behavioral approaches represents a significant advancement in responsible data science research, suggesting that future work should continue to address both dimensions rather than treating them as separate domains.

Future research could further explore the dynamic interplay between technical solutions and human factors (as introduced in section 3.1), recognizing that these elements function as an integrated sociotechnical system rather than isolated components. While this dissertation has separated these aspects for conceptual clarity, practical implementations must account for their inherent interdependence. Promising research directions include developing feedback mechanisms where technical systems adapt to observed human behavioral patterns, as well as investigating how technical design choices influence practitioner behavior and decision-making processes.

7.2 Theoretical Translation Across Disciplines as a Methodological Innovation

A significant contribution of this dissertation is the methodological innovation of translating theoretical frameworks across seemingly disparate disciplines. By applying behavior change theories from cognitive and clinical psychology to data science practices, I have demonstrated the value of cross-disciplinary theoretical adaptation. This translation required careful consideration of contextual differences and domain-specific challenges, yet yielded valuable insights that would have been difficult to achieve within the confines of traditional data science research. This methodological approach has broader implications for responsible technology research, suggesting that similar translations might prove valuable for addressing ethical challenges in adjacent fields such as software engineering, artificial intelligence, and human-computer

interaction. Future work should continue to explore such cross-disciplinary theoretical translations as a means of developing novel solutions to complex sociotechnical problems.

7.3 Bridging Theory and Practice in Responsible Data Science

The 5W1H design space represents a crucial bridge between theoretical frameworks and practical implementation of responsible data science. While behavior change theories provide valuable conceptual foundations (as discussed in chapter 3), translating these theories into actionable interventions requires structured guidance that is domain-specific. Our interrogative framework demonstrates how established psychological theories can be operationalized within data science workflows by prompting designers to consider the full contextual dimensions of an intervention. This translation process highlights the importance of interdisciplinary approaches to responsible data science—combining knowledge from psychology, human-computer interaction, and data science to create effective solutions. Future work should continue to explore this translation process, potentially incorporating emerging theories from related fields like decision science and organizational behavior to further enrich the design space.

7.4 Balancing Individual and Systemic Approaches to Change

The development of our design space reveals a fundamental tension in responsible data science interventions: the balance between targeting individual behavior change and addressing broader systemic factors. While our framework primarily focuses on

interventions at the individual practitioner level, the interrogative structure (particularly through the "Who" and "Why" dimensions) acknowledges the importance of organizational and societal contexts. This tension echoes the broader conceptual framework introduced in chapter 3, which positions responsible data science at the intersection of human factors and technical systems. As the field matures, intervention designers must increasingly consider multi-level approaches that simultaneously address individual behaviors, team dynamics, organizational policies, and industry standards. Our design space provides a starting point for this integrated approach, but future work should explore how to effectively coordinate interventions across these different levels to create sustainable change in responsible data science practices.

7.5 The Role of Visualization in Promoting Responsible Practices

Our work with PreFair demonstrates the significant potential of visualization tools as behavior change interventions in responsible data science. By making the complex relationships between pre-processing choices and fairness outcomes visible and explorable, PreFair transforms abstract fairness concerns into concrete, actionable insights. This approach addresses a critical gap in existing responsible data science practices: the difficulty practitioners face in understanding the downstream consequences of their decisions. The visual presentation of fairness trade-offs shifts this cognitive burden from mental modeling to direct perception, potentially reducing the barriers to responsible practice. This finding suggests that future behavior change interventions could leverage visualization techniques to make ethical considerations more accessible and actionable throughout the data science workflow. Rather than treating ethics as separate from technical practice, tools like PreFair demonstrate how ethical considerations can be seamlessly integrated into practitioners' decision-making

processes, potentially leading to more consistent application of responsible practices. Future research could investigate interactive visual tools that illustrate how different choices among the entire data science stages could affect multiple ethical dimensions simultaneously, enabling practitioners to navigate complex multi-dimensional ethical spaces more effectively.

7.6 Intervention Design Should Balance Technical Capability with Workflow Integration

The development of PreFair revealed a crucial tension in behavior change intervention design: the balance between technical sophistication and workflow integration. While PreFair successfully provided powerful capabilities for fairness analysis, its effectiveness was limited by challenges in integrating with existing data science workflows. This highlights a fundamental principle for intervention design: tools should not only provide technical capabilities for responsible practice but should also align with practitioners' established work patterns to achieve adoption and impact. Our experience suggests that effective interventions should be designed with a deep understanding of practitioners' existing workflows, constraints, and priorities. This finding expands our understanding of the "opportunity" component in our behavior change framework, emphasizing that interventions must create favorable conditions for adoption within existing professional contexts. Future work should prioritize user-centered design approaches that incorporate workflow analysis as a foundational step in intervention development, ensuring that tools for responsible data science enhance rather than disrupt practitioners' ability to accomplish their primary goals.

7.7 The Tension Between Intervention Efficacy and Cognitive Load

This research revealed a fundamental tension in responsible data science interventions: tools that produce the strongest fairness improvements may also impose the highest cognitive demands. The Aequitas intervention significantly enhanced fairness metrics and prompted more responsible behaviors, but at the cost of increased mental demand, physical demand, and effort. This tension highlights a critical design challenge for the field—creating interventions that effectively guide responsible practice without overwhelming practitioners' cognitive resources. Our findings suggest that optimal intervention design may involve strategic deployment: using comprehensive tools like Aequitas at critical decision points while employing lightweight approaches for day-to-day work. Furthermore, the acceptability of cognitive load appears context-dependent, influenced by factors such as the stakes of algorithmic decisions, organizational priorities, and practitioners' personal connection to fairness concerns. To the end of maximizing the interventions' efficacy while maintaining the cognitive load it brings to the users, future research should investigate how intelligent assistance and adaptive interfaces might dynamically manage cognitive load while maintaining intervention effectiveness. Beyond this potential research direction that seek to balance the intervention efficacy and cognitive load, studies should explore the relationship between cognitive load and intervention sustainability, examining whether high-demand tools lead to abandonment or workarounds over time, and identifying thresholds of acceptable cognitive burden across different practitioner contexts and experience levels.

7.8 Bridging the Gap Between Behavioral Change and Outcome Improvement

My thesis reveals an important distinction between promoting responsible behaviors and improving fairness outcomes. While I observed significant increases in responsible behaviors with both interventions, these behavioral changes translated unevenly to fairness improvements. This disconnect challenges a key assumption in responsible data science: that following responsible practices automatically ensures fair outcomes. Future research should investigate this relationship more deeply by (1) identifying which specific behaviors most reliably predict improved fairness metrics; (2) developing interventions that strategically target these high-impact behaviors rather than broadly promoting all responsible practices; and (3) creating evaluation frameworks that more comprehensively track both the adoption of behaviors, ultimate model fairness changes and building connections between these behaviors with their downstream effects on model fairness. Understanding this behavior-outcome relationship could enable more efficient interventions that maximize fairness improvements while minimizing the cognitive burden placed on practitioners.

Chapter 8

Conclusions

In summary, this dissertation addressed three key research questions that collectively advance responsible data science through behavioral interventions. In response to **RQ1** (*How can I utilize behavior change theories to inform a novel framework for responsible data science using the lens of behavior change interventions?*), I developed a comprehensive theoretical framework that adapts established behavior change theories to the data science context, demonstrating how capability, opportunity, and motivation factors collectively influence responsible practices. I found that behavior change theories can inform responsible data science by providing a structured framework that addresses capability, opportunity, and motivation as essential components for transforming technical knowledge into responsible practice.

In response to **RQ2** (*How can I scaffold the design and development of behavior change interventions for responsible data science?*), I synthesized a design space for behavior change interventions that provides structured guidance for intervention designers, highlighting key variables for consideration across behavioral and implementation dimensions. I found that effective scaffolding for behavior change interventions in responsible data science requires a comprehensive design space that accounts for factors like temporal dimensions, intervention mechanisms, and integration points

within existing data science workflows.

Finally, addressing **RQ3** (*How effective are behavior change interventions at improving responsible data science outcomes?*), my work with PreFair and subsequent intervention efficacy evaluation studies revealed that effective behavior change interventions should balance concrete ethical metrics with seamless workflow integration, while comprehensive evaluation measurements are essential for measuring intervention efficacy across both behaviors and outcomes. I found that behavior change interventions are most effective when they combine concrete metrics that make ethical considerations visible with seamless workflow integration, while requiring structured evaluation frameworks to measure their impact on data scientists' decision-making processes. These contributions collectively advance the state-of-the-art in promoting responsible data science through behavior change interventions.

Bibliography

[1] The trifacta data engineering cloud, Aug 2012. URL https://www.trifacta.com/.

[2] South German Credit. UCI Machine Learning Repository, 2020. DOI: https://doi.org/10.24432/C5QG88.

[3] Charles Abraham and Susan Michie. A taxonomy of behavior change techniques used in interventions. *Health Psychology*, 27(3):379–387, 2008. ISSN 1930-7810, 0278-6133. doi: 10.1037/0278-6133.27.3.379. URL http://doi.apa.org/getdoi.cfm?doi=10.1037/0278-6133.27.3.379.

[4] Accountability Act. Health insurance portability and accountability act of 1996. *Public law*, 104:191, 1996.

[5] Philip Adler, Casey Falk, Sorelle A Friedler, Tionney Nix, Gabriel Rybeck, Carlos Scheidegger, Brandon Smith, and Suresh Venkatasubramanian. Auditing black-box models for indirect influence. *Knowledge and Information Systems*, 54:95–122, 2018.

[6] Evidently AI. Evidently: Evaluate and monitor ML models from validation to production. Evidently AI, 2022. URL https://github.com/evidentlyai/evidently.

[7] Ifeoma Ajunwa and Daniel Greene. Platforms at work: Automated hiring plat-

forms and other new intermediaries in the organization of work. *CGN: Sociology (Topic)*, 2019.

[8] Zaher Ali Al-Sai, Rosni Abdullah, and Mohd Heikal Husin. Critical success factors for big data: a systematic literature review. *IEEE Access*, 8:118940–118956, 2020.

[9] Saleema Amershi, Andrew Begel, Christian Bird, Robert DeLine, Harald Gall, Ece Kamar, Nachiappan Nagappan, Besmira Nushi, and Thomas Zimmermann. Software Engineering for Machine Learning: A Case Study. In *2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*, pages 291–300, Montreal, QC, Canada, May 2019. IEEE. ISBN 978-1-72811-760-7. doi: 10.1109/ICSE-SEIP.2019.00042. URL `https://ieeexplore.ieee.org/document/8804457/`.

[10] Mihael Ankerst, Markus M Breunig, Hans-Peter Kriegel, and Jörg Sander. Optics: Ordering points to identify the clustering structure. *ACM Sigmod record*, 28(2):49–60, 1999.

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

[12] Lou Atkins, Jill Francis, Rafat Islam, Denise Oâ€™Connor, Andrea Patey, Noah Ivers, Robbie Foy, Eilidh M. Duncan, Heather Colquhoun, Jeremy M. Grimshaw, Rebecca Lawton, and Susan Michie. A guide to using the Theoretical Domains Framework of behaviour change to investigate implementation problems. *Implementation Science*, 12(1):77, December 2017. ISSN 1748-5908. doi: 10.1186/s13012-017-0605-9. URL `http://implementationscience.biomedcentral.com/articles/10.1186/s13012-017-0605-9`.

[13] Teresa K Attwood, Sarah Blackford, Michelle D Brazas, Angela Davies, and

Maria Victoria Schneider. A global perspective on evolving bioinformatics and data science training needs. *Briefings in Bioinformatics*, 20(2):398–404, 2019.

[14] Alan Baddeley. *Working memory, thought, and action*, volume 45. OuP Oxford, 2007.

[15] Chelsea Barabas, Madars Virza, Karthik Dinakar, Joichi Ito, and Jonathan Zittrain. Interventions over predictions: Reframing the ethical debate for actuarial risk assessment. In Sorelle A. Friedler and Christo Wilson, editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 62–76. PMLR, 23–24 Feb 2018. URL `https://proceedings.mlr.press/v81/barabas18a.html`.

[16] Solon Barocas and Danah Boyd. Engaging the ethics of data science in practice. *Communications of the ACM*, 60:23 – 25, 2017.

[17] Solon Barocas and Andrew D. Selbst. Big Data's Disparate Impact. *SSRN eLibrary*, 2014.

[18] Rachel Bellamy, Kuntal Dey, Michael Hind, Samuel Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush Varshney, and Yunfeng Zhang. Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, PP:1–1, 09 2019. doi: 10.1147/JRD.2019.2942287.

[19] Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilović, et al. Ai fairness 360: An extensible toolkit for detecting

and mitigating algorithmic bias. *IBM Journal of Research and Development*, 63(4/5):4–1, 2019.

[20] Vaishnavi Bhargava, Miguel Couceiro, and Amedeo Napoli. Limeout: an ensemble approach to improve process fairness. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 475–491. Springer, 2020.

[21] Avinash Bhat, Austin Coursey, Grace Hu, Sixian Li, Nadia Nahar, Shurui Zhou, Christian Kästner, and Jin L. C. Guo. Aspirations and Practice of Model Documentation: Moving the Needle with Nudging and Traceability. In *CHI*, 2023. doi: 10.1145/3544548.3581518. URL `http://arxiv.org/abs/2204.06425`.

[22] Carsten Binnig, Lorenzo De Stefani, Tim Kraska, Eli Upfal, Emanuel Zgraggen, and Zheguang Zhao. Toward sustainable insights, or why polygamy is bad for you. In *CIDR*, 2017.

[23] Derya Birant and Alp Kut. St-dbscan: An algorithm for clustering spatial–temporal data. *Data & knowledge engineering*, 60(1):208–221, 2007.

[24] Sumon Biswas and Hridesh Rajan. Fair preprocessing: towards understanding compositional fairness of data transformers in machine learning pipeline. *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, 2021.

[25] David M. Blei and Padhraic Smyth. Science and data science. *Proceedings of the National Academy of Sciences*, 114(33):8689–8692, 2017. doi: 10.1073/pnas.1702076114. URL `https://www.pnas.org/doi/abs/10.1073/pnas.1702076114`.

[26] Ashley Boone, Carl Disalvo, and Christopher A Le Dantec. Data practice for a politics of care: Food assistance as a site of careful data work. In *Proceedings*

of the 2023 CHI Conference on Human Factors in Computing Systems, pages 1–13, 2023.

[27] Belinda Borrelli and Robin Mermelstein. Goal setting and behavior change in a smoking cessation program. *Cognitive Therapy and Research*, 18(1):69–83, 1994.

[28] Jason Brownlee. *Data preparation for machine learning: data cleaning, feature selection, and data transforms in Python*. Machine Learning Mastery, 2020.

[29] Robert J Brunner and Edward J Kim. Teaching data science. *Procedia Computer Science*, 80:1947–1956, 2016.

[30] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018.

[31] Steven P. Callahan, Juliana Freire, Emanuele Santos, Carlos E. Scheidegger, Cláudio T. Silva, and Huy T. Vo. Vistrails: Visualization meets data management. In *In ACM SIGMOD*, pages 745–747. ACM Press, 2006.

[32] Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. Optimized pre-processing for discrimination prevention. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL `https://proceedings.neurips.cc/paper/2017/file/9a49a25d845a483fae4be7e341368e36-Paper.pdf`.

[33] James Cane, Denise Oâ€™Connor, and Susan Michie. Validation of the theoretical domains framework for use in behaviour change and implementa-

tion research. *Implementation Science*, 7(1):37, December 2012. ISSN 1748-5908. doi: 10.1186/1748-5908-7-37. URL http://implementationscience.biomedcentral.com/articles/10.1186/1748-5908-7-37.

[34] Rachel N Carey, Lauren E Connell, Marie Johnston, Alexander J Rothman, Marijn de Bruin, Michael P Kelly, and Susan Michie. Behavior Change Techniques and Their Mechanisms of Action: A Synthesis of Links Described in Published Intervention Literature. *Annals of Behavioral Medicine*, October 2018. ISSN 0883-6612, 1532-4796. doi: 10.1093/abm/kay078. URL https://academic.oup.com/abm/advance-article/doi/10.1093/abm/kay078/5126198.

[35] Rachel N Carey, Lauren E Connell, Marie Johnston, Alexander J Rothman, Marijn De Bruin, Michael P Kelly, and Susan Michie. Behavior change techniques and their mechanisms of action: a synthesis of links described in published intervention literature. *Annals of Behavioral Medicine*, 53(8):693–707, 2019.

[36] Archie B Carroll. The pyramid of corporate social responsibility: Toward the moral management of organizational stakeholders. *Business horizons*, 34(4): 39–48, 1991.

[37] Sun Ju Chang, Suyoung Choi, Se-An Kim, and Misoon Song. Intervention strategies based on information-motivation-behavioral skills model for health behavior change: a systematic review. *Asian Nursing Research*, 8(3):172–181, 2014.

[38] Saurabh Chaudhari, Suparna Ghanvatkar, Atreyi Kankanhalli, et al. Personalization of intervention timing for physical activity: scoping review. *JMIR mHealth and uHealth*, 10(2):e31327, 2022.

[39] Lu Cheng, Kush R Varshney, and Huan Liu. Socially responsible ai algorithms: Issues, purposes, and challenges. *Journal of Artificial Intelligence Research*, 71: 1137–1181, 2021.

[40] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5, 10 2016. doi: 10.1089/big.2016.0047.

[41] Edgar F Codd. A relational model of data for large shared data banks. In *Software pioneers*, pages 263–294. Springer, 2002.

[42] Michael Correll. Ethical dimensions of visualization research. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–13, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450359702. doi: 10.1145/3290605.3300418. URL `https://doi.org/10.1145/3290605.3300418`.

[43] Michael Correll and Michael Gleicher. Error Bars Considered Harmful: Exploring Alternate Encodings for Mean and Error. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2142–2151, December 2014. ISSN 1077-2626. doi: 10.1109/TVCG.2014.2346298. URL `http://ieeexplore.ieee.org/document/6875915/`.

[44] Henriette Cramer, Jean Garcia-Gathright, Sravana Reddy, Aaron Springer, and Romain Takeo Bouyer. Translation, tracks & data: An algorithmic bias effort in practice. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI EA '19, page 1–8, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450359719. doi: 10.1145/3290607.3299057. URL `https://doi.org/10.1145/3290607.3299057`.

[45] Anamaria Crisan and Brittany Fiore-Gartland. Fits and starts: Enterprise

use of automl and the role of humans in the loop. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380966. doi: 10.1145/3411764.3445775. URL https://doi.org/10.1145/3411764.3445775.

[46] Anamaria Crisan, Brittany Fiore-Gartland, and Melanie Tory. Passing the data baton: A retrospective analysis on data science work and workers. *IEEE Transactions on Visualization and Computer Graphics*, PP:1–1, 10 2020. doi: 10.1109/TVCG.2020.3030340.

[47] Anamaria Crisan, Brittany Fiore-Gartland, and Melanie Tory. Passing the data baton : A retrospective analysis on data science work and workers. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1860–1870, 2021. doi: 10.1109/TVCG.2020.3030340.

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

[49] Wesley Hanwen Deng, Manish Nagireddy, Michelle Seng Ah Lee, Jatinder Singh, Zhiwei Steven Wu, Kenneth Holstein, and Haiyi Zhu. Exploring how machine learning practitioners (try to) use fairness toolkits. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 473–484, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533113. URL https://doi.org/10.1145/3531146.3533113.

[50] Pietro G Di Stefano, James M Hickey, and Vlasios Vasileiou. Counterfac-

tual fairness: removing direct effects through regularization. *arXiv preprint arXiv:2002.10774*, 2020.

[51] Jie Ding, Vahid Tarokh, and Yuhong Yang. Model selection techniques: An overview. *IEEE Signal Processing Magazine*, 35(6):16–34, 2018.

[52] Yiran Dong and Chao-Ying Joanne Peng. Principled missing data methods for researchers. *SpringerPlus*, 2:1–17, 2013.

[53] Ziwei Dong, Teanna Barrett, Ameya B Patil, Yuichi Shoda, Leilani Battle, and Emily Wall. A design space of behavior change interventions for responsible data science. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*, IUI '25, page 37–53, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400713064. doi: 10.1145/3708359.3712140. URL https://doi.org/10.1145/3708359.3712140.

[54] Ziwei Dong, Fang Cao, Eli T Brown, and Emily Wall. Fairness first: A visual analytic approach to exploring fairness in pre-processing permutations. 2025. in preparation.

[55] Ziwei Dong, Ameya Patil, Yuichi Shoda, Leilani Battle, and Emily Wall. Behavior matters: An alternative perspective on promoting responsible data science. *Proceedings of the ACM Conference on Computer Supported Cooperative Work & Social Computing*, 2025.

[56] Ziwei Dong, Keke Wu, Leilani Battle, and Emily Wall. Evaluating the efficacy of behavior change interventions for responsible data science. 2025. in preparation.

[57] David Donoho. 50 years of data science. *Journal of Computational and Graphical Statistics*, 26(4):745–766, 2017. doi: 10.1080/10618600.2017.1384734. URL https://doi.org/10.1080/10618600.2017.1384734.

[58] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

[59] David A Dzewaltowski, Paul A Estabrooks, Lisa M Klesges, Sheana Bull, and Russell E Glasgow. Behavior change intervention research in community settings: how generalizable are the results? *Health promotion international*, 19 (2):235–245, 2004.

[60] Upol Ehsan, Philipp Wintersberger, Elizabeth Anne Watkins, Carina Manger, Gonzalo Ramos, Justin Weisz, Hal Daumé Iii, Andreas Riener, and Mark Riedl. Human-centered explainable ai: Coming of age. In *ACM CHI Conference on Human Factors in Computing Systems*, 2023.

[61] Virginia Eubanks. *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press, 2018.

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

[63] Mi Feng, Cheng Deng, Evan M. Peck, and Lane Harrison. Hindsight: Encouraging exploration through direct encoding of personal interaction history. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):351–360, 2017. doi: 10.1109/TVCG.2016.2599058.

[64] U.M. Feyyad. Data mining and knowledge discovery: making sense out of data. *IEEE Expert*, 11(5):20–25, 1996. doi: 10.1109/64.539013.

[65] U.M. Feyyad. Data mining and knowledge discovery: making sense out of data. *IEEE Expert*, 11(5):20–25, 1996. doi: 10.1109/64.539013.

[66] Martin Fishbein and Icek Ajzen. Theory-based behavior change interventions: Comments on hobbis and sutton. *Journal of health psychology*, 10(1):27–31, 2005.

[67] Martin Fishbein, Triandis Hc, Kanfer Fh, Marshall H. Becker, and Susan E. Middlestadt. Factors influencing behavior and behavior change. 2000.

[68] Brianna S Fjeldsoe, Alison L Marshall, and Yvette D Miller. Behavior change interventions delivered by mobile telephone short-message service. *American journal of preventive medicine*, 36(2):165–173, 2009.

[69] Anthony W. Flores, K. Bechtel, and Christopher T. Lowenkamp. False positives, false negatives, and false analyses: A rejoinder to "machine bias: There's software used across the country to predict future criminals. and it's biased against blacks". *Federal Probation*, 80:38, 2016.

[70] Imola K Fodor. A survey of dimension reduction techniques. Technical report, Lawrence Livermore National Lab., CA (US), 2002.

[71] James Fogarty, Scott E Hudson, Christopher G Atkeson, Daniel Avrahami, Jodi Forlizzi, Sara Kiesler, Johnny C Lee, and Jie Yang. Predicting human interruptibility with sensors. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 12(1):119–146, 2005.

[72] BJ Fogg. A behavior model for persuasive design. In *Proceedings of the 4th International Conference on Persuasive Technology*, Persuasive '09, New York, NY, USA, 2009. Association for Computing Machinery. ISBN 9781605583761. doi: 10.1145/1541948.1541999. URL `https://doi.org/10.1145/1541948.1541999`.

[73] Brian J Fogg. A behavior model for persuasive design. In *Proceedings of the 4th international Conference on Persuasive Technology*, pages 1–7, 2009.

[74] Sorelle Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. 12 2014. doi: 10.1145/2783258.2783311.

[75] Ursula Garzcarek and Detlef Steuer. Approaching ethical guidelines for data scientists. *Applications in statistical computing: From music data analysis to industrial quality improvement*, pages 151–169, 2019.

[76] Golnaz Ghiasi, Tsung-Yi Lin, and Quoc V Le. Dropblock: A regularization method for convolutional networks. *Advances in neural information processing systems*, 31, 2018.

[77] Lisa Gitelman. *"Raw data" is an oxymoron*. The MIT Press, 2013.

[78] Usman Gohar, Zeyu Tang, Jialu Wang, Kun Zhang, Peter Spirtes, Yang Liu, and Lu Cheng. Long-term fairness inquiries and pursuits in machine learning: A survey of notions, methods, and challenges, 06 2024.

[79] Thomas S Griffith and Thomas A Ferguson. Cell death in the maintenance and abrogation of tolerance: the five ws of dying cells. *Immunity*, 35(4):456–466, 2011.

[80] Grace Guo, Ehud Karavani, Alex Endert, and Bum Chul Kwon. Causalvis: Visualizations for Causal Inference. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023. doi: 10.1145/3544548.3581236.

[81] Brian D Hall, Yang Liu, Yvonne Jansen, Pierre Dragicevic, Fanny Chevalier, and Matthew Kay. A survey of tasks and visualizations in multiverse analysis reports. In *Computer Graphics Forum*, volume 41, pages 402–426. Wiley Online Library, 2022.

[82] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in super-

vised learning. *Advances in neural information processing systems*, 29:3315–3323, 2016.

[83] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

[84] Galen Harrison, Kevin Bryson, Ahmad Emmanuel Balla Bamba, Luca Dovichi, Aleksander Herrmann Binion, Arthur Borem, and Blase Ur. Jupyterlab in retrograde: Contextual notifications that highlight fairness and bias issues for data scientists. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–19, 2024.

[85] Geoff Hart. The five w's of online help systems. *Geoff Hart*, 2002.

[86] Sandra G Hart. Nasa-task load index (nasa-tlx); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting*, volume 50, pages 904–908. Sage publications Sage CA: Los Angeles, CA, 2006.

[87] Julian Heinrich and Daniel Weiskopf. State of the art of parallel coordinates. *Eurographics (state of the art reports)*, pages 95–116, 2013.

[88] Anne Hove Henriksen Heise, Soraj Hongladarom, Anna Jobin, Katharina Kinder-Kurlanda, Sun Sun, Elisabetta Locatelli Lim, Annette Markham, Paul J Reilly, Katrin Tiidenberg, and Carsten Wilhelm. Internet research: Ethical guidelines 3.0. 2019.

[89] Sylvia C Hewitt. The five w's of progesterone receptors a and b: now we know where and when. *Endocrinology*, 147(12):5501–5502, 2006.

[90] Simon David Hirsbrunner, Michael Tebbe, and Claudia Mueller-Birn. From critical technical practice to reflexive data science. *CONVERGENCE-THE INTERNATIONAL JOURNAL OF RESEARCH INTO NEW MEDIA TECH-*

NOLOGIES, 30(1):190–215, FEB 2024. ISSN 1354-8565. doi: 10.1177/13548565221132243.

[91] Fred Hohman, Minsuk Kahng, Robert Pienta, and Duen Horng Chau. Visual analytics in deep learning: An interrogative survey for the next frontiers. *IEEE transactions on visualization and computer graphics*, 25(8):2674–2693, 2018.

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

[93] Eric Horvitz. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '99, page 159–166, New York, NY, USA, 1999. Association for Computing Machinery. ISBN 0201485591. doi: 10.1145/302979.303030. URL `https://doi.org/10.1145/302979.303030`.

[94] Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933.

[95] IBM. watsonx.governance. IBM, 2020. URL `https://www.ibm.com/products/watsonx-governance`.

[96] Vasileios Iosifidis, Besnik Fetahu, and Eirini Ntoutsi. Fae: A fairness-aware ensemble framework. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 1375–1380. IEEE, 2019.

[97] Sérgio Jesus, Pedro Saleiro, Beatriz M Jorge, Rita P Ribeiro, João Gama, Pedro Bizarro, Rayid Ghani, et al. Aequitas flow: Streamlining fair ml experimentation. *arXiv preprint arXiv:2405.05809*, 2024.

[98] Anna Jobin, Marcello Ienca, and Effy Vayena. The global landscape of ai ethics guidelines. *Nature Machine Intelligence*, 1, 09 2019. doi: 10.1038/s42256-019-0088-2.

[99] Eunice Jun, Maureen Daum, Jared Roesch, Sarah Chasins, Emery Berger, Rene Just, and Katharina Reinecke. Tea: A high-level language and runtime system for automating statistical analysis. pages 591–603, 10 2019. ISBN 978-1-4503-6816-2. doi: 10.1145/3332165.3347940.

[100] Daniel Kahneman. *Thinking Fast and Slow*. Macmillan Publishers, 2011.

[101] S KailashKarthik. The impossibility theorem of machine fairness - a causal perspective. *ArXiv*, abs/2007.06024, 2020.

[102] Faisal Kamiran and Toon Calders. Data pre-processing techniques for classification without discrimination. *Knowledge and Information Systems*, 33, 10 2011. doi: 10.1007/s10115-011-0463-8.

[103] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Fairness-aware classifier with prejudice remover regularizer. In *ECML/PKDD*, 2012. URL `https://api.semanticscholar.org/CorpusID:10352172`.

[104] Shubhra Kanti Karmaker ("Santu"), Md. Mahadi Hassan, Micah J. Smith, Lei Xu, Chengxiang Zhai, and Kalyan Veeramachaneni. Automl to date and beyond: Challenges and opportunities. *ACM Comput. Surv.*, 54(8), oct 2021. ISSN 0360-0300. doi: 10.1145/3470918. URL `https://doi.org/10.1145/3470918`.

[105] Herbert C Kelman. Compliance, identification, and internalization three processes of attitude change. *Journal of conflict resolution*, 2(1):51–60, 1958.

[106] John T Kent. Information gain and a general measure of correlation. *Biometrika*, 70(1):163–173, 1983.

[107] Yones Khaledian and Bradley A Miller. Selecting appropriate machine learning methods for digital soil mapping. *Applied Mathematical Modelling*, 81:401–418, 2020.

[108] Thomas Kluyver, Benjamin Ragan-Kelley, Fernando Pérez, Brian Granger, Matthias Bussonnier, Jonathan Frederic, Kyle Kelley, Jessica Hamrick, Jason Grout, Sylvain Corlay, et al. Jupyter notebooks–a publishing format for reproducible computational workflows. In *Positioning and power in academic publishing: Players, agents and agendas*, pages 87–90. IOS press, 2016.

[109] Phillippa Lally and Benjamin Gardner. Promoting habit formation. *Health psychology review*, 7(sup1):S137–S158, 2013.

[110] Michelle S. Lam, Zixian Ma, Anne Li, Izequiel Freitas, Dakuo Wang, James A. Landay, and Michael S. Bernstein. Model Sketching: Centering Concepts in Early-Stage Machine Learning Model Design. *arXiv 2303.02884*, 2023. doi: 10.1145/3544548.3581290. URL `http://arxiv.org/abs/2303.02884`.

[111] Elsie Lee-Robbins and Eytan Adar. Affective Learning Objectives for Communicative Visualizations. *IEEE Transactions on Visualization and Computer Graphics*, pages 1–11, 2022. ISSN 1077-2626, 1941-0506, 2160-9306. doi: 10.1109/TVCG.2022.3209500. URL `https://ieeexplore.ieee.org/document/9905872/`.

[112] Haotian Li, Lu Ying, Haidong Zhang, Yingcai Wu, Huamin Qu, and Yun Wang. Notable: On-the-fly Assistant for Data Storytelling in Computational Notebooks. In *CHI*, 2023. doi: 10.1145/3544548.3580965. URL `https://dl.acm.org/doi/10.1145/3544548.3580965`.

[113] Xingjun Li, Yizhi Zhang, Justin Leung, Chengnian Sun, and Jian Zhao. EDAssistant: Supporting Exploratory Data Analysis in Computational Notebooks with In Situ Code Search and Recommendation. *ACM TiiS*, 13, 2023. doi: 10.1145/3545995. URL `https://dl.acm.org/doi/10.1145/3545995`.

[114] Yanhui Li, Linghan Meng, Lin Chen, Li Yu, Di Wu, Yuming Zhou, and Baowen Xu. Training data debugging for the fairness of machine learning software. In *Proceedings of the 44th International Conference on Software Engineering*, pages 2215–2227, 2022.

[115] Alexander V Lotov and Kaisa Miettinen. Visualizing the pareto frontier. *Multiobjective optimization*, 5252:213–243, 2008.

[116] Sabrigiriraj M and K. Manoharan. Teaching machine learning and deep learning introduction: An innovative tutorial-based practical approach. *WSEAS TRANSACTIONS ON ADVANCES in ENGINEERING EDUCATION*, 21:54–61, 06 2024. doi: 10.37394/232010.2024.21.8.

[117] Michael A. Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. Co-designing checklists to understand organizational challenges and opportunities around fairness in ai. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–14, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450367080. doi: 10.1145/3313831.3376445. URL `https://doi.org/10.1145/3313831.3376445`.

[118] Harry D. Mafukidze, Action Nechibvute, Abid Yahya, Irfan Anjum Badruddin, Sarfaraz Kamangar, and Mohamed Hussien. Development of a modularized undergraduate data science and big data curricular using no-code software development tools. *IEEE ACCESS*, 12:100939–100956, 2024. ISSN 2169-3536. doi: 10.1109/ACCESS.2024.3429241.

[119] Ali A. Mahmoud, Tahani AL Shawabkeh, Walid A. Salameh, and Ibrahim Al Amro. Performance predicting in hiring process and performance appraisals using machine learning. In *2019 10th International Conference on Information and Communication Systems (ICICS)*, pages 110–115, 2019. doi: 10.1109/IACS.2019.8809154.

[120] Serena Mastria, Alessandro Vezzil, and Andrea De Cesarei. Going green: A review on the role of motivation in sustainable behavior. *Sustainability*, 15(21): 15429, 2023.

[121] Mary L McHugh. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282, 2012.

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

[123] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, 54(6), jul 2021. ISSN 0360-0300. doi: 10.1145/3457607. URL `https://doi.org/10.1145/3457607`.

[124] Alana I Mendelsohn. Creatures of habit: The neuroscience of habit and purposeful behavior. *Biological psychiatry*, 85(11):e49–e51, 2019.

[125] Amanda Meng, Carl DiSalvo, and Ellen Zegura. Collaborative data work towards a caring democracy. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–23, 2019.

[126] Aditya Krishna Menon and Robert C Williamson. The cost of fairness in classification. *arXiv preprint arXiv:1705.09055*, 2017.

[127] S Michie. Making psychological theory useful for implementing evidence based practice: a consensus approach. *Quality and Safety in Health Care*, 14(1): 26–33, February 2005. ISSN 1475-3898, 1475-3901. doi: 10.1136/qshc.2004.011155. URL https://qualitysafety.bmj.com/lookup/doi/10.1136/qshc.2004.011155.

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

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

[130] Susan Michie, Michelle Richardson, Marie Johnston, Charles Abraham, Jill Francis, Wendy Hardeman, Martin P. Eccles, James Cane, and Caroline E. Wood. The Behavior Change Technique Taxonomy (v1) of 93 Hierarchically Clustered Techniques: Building an International Consensus for the Reporting of Behavior Change Interventions. *Annals of Behavioral Medicine*, 46(1):81–95, August 2013. ISSN 0883-6612, 1532-4796. doi: 10.1007/s12160-013-9486-6. URL https://academic.oup.com/abm/article/46/1/81-95/4563254.

[131] Microsoft. Interpret Community SDK, 2019. URL https://github.com/interpretml/interpret-community.

[132] Microsoft. Responsible AI Toolbox. Microsoft, 2020. URL https://github.com/microsoft/responsible-ai-toolbox.

[133] Carla K Miller. Adaptive intervention designs to promote behavioral change in adults: what is the evidence? *Current diabetes reports*, 19:1–9, 2019.

[134] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 220–229, 2019.

[135] Molly Mulshine. A major flaw in google's algorithm allegedly tagged two black people's faces with the word'gorillas'. *Business Insider*, 2015.

[136] David Munechika, Zijie J. Wang, Jack Reidy, Josh Rubin, Krishna Gade, Krishnaram Kenthapadi, and Duen Horng Chau. Visual Auditor: Interactive Visualization for Detection and Summarization of Model Biases. In *VIS*, 2022. doi: 10.1109/VIS54862.2022.00018.

[137] Kristian S. Nielsen, Sander van der Linden, and Paul C. Stern. How Behavioral Interventions Can Reduce the Climate Impact of Energy Use. *Joule*, 4 (8):1613–1616, 2020. ISSN 2542-4351. doi: https://doi.org/10.1016/j.joule.2020.07.008. URL `https://www.sciencedirect.com/science/article/pii/S2542435120303263`.

[138] Harsha Nori, Samuel Jenkins, Paul Koch, and Rich Caruana. InterpretML: A Unified Framework for Machine Learning Interpretability. *arXiv*, 2019. URL `http://arxiv.org/abs/1909.09223`.

[139] Rita Orji and Karyn Moffatt. Persuasive technology for health and wellness: State-of-the-art and emerging trends. *Health informatics journal*, 24(1):66–91, 2018.

[140] Giuliano Orru and Luca Longo. The evolution of cognitive load theory and the measurement of its intrinsic, extraneous and germane loads: a review. In *Human Mental Workload: Models and Applications: Second International Symposium,*

H-WORKLOAD 2018, Amsterdam, The Netherlands, September 20-21, 2018, Revised Selected Papers 2, pages 23–48. Springer, 2019.

[141] Jeni Paay, Jesper Kjeldskov, Mikael B. Skov, Lars Lichon, and Stephan Rasmussen. Understanding individual differences for tailored smoking cessation apps. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, page 1699–1708, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450331456. doi: 10.1145/2702123.2702321. URL https://doi.org/10.1145/2702123.2702321.

[142] R. Parasuraman, T.B. Sheridan, and C.D. Wickens. A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 30(3):286–297, 2000. doi: 10.1109/3468.844354.

[143] Charlie Pinder, Jo Vermeulen, Benjamin R. Cowan, and Russell Beale. Digital behaviour change interventions to break and form habits. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 25:1 – 66, 2018.

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

[145] Robert W Proctor and Darryl W Schneider. Hick's law for choice reaction time: A review. *Quarterly Journal of Experimental Psychology*, 71(6):1281–1299, 2018.

[146] Formerly Data Protection. General data protection regulation (gdpr). *Intersoft Consulting, Accessed in October*, 24(1), 2018.

[147] Ross Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1993.

[148] Henriette Rau, Susanne Nicolai, and Susanne Stoll-Kleemann. A systematic review to assess the evidence-based effectiveness, content, and success factors of behavior change interventions for enhancing pro-environmental behavior in individuals. *Frontiers in Psychology*, 13, 2022. ISSN 1664-1078. doi: 10.3389/fpsyg.2022.901927. URL https://www.frontiersin.org/articles/10.3389/fpsyg.2022.901927.

[149] Shaina Raza, Shardul Ghuge, Chen Ding, Elham Dolatabadi, and Deval Pandya. Fair enough: Develop and assess a fair-compliant dataset for large language model training? *Data Intelligence*, 6(2):559–585, 2024.

[150] Anna Rogers, Tim Baldwin, and Kobi Leins. Just what do you think you're doing, dave?'a checklist for responsible data use in nlp. *arXiv preprint arXiv:2109.06598*, 2021.

[151] Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987. ISSN 0377-0427. doi: https://doi.org/10.1016/0377-0427(87)90125-7. URL https://www.sciencedirect.com/science/article/pii/0377042787901257.

[152] Pedro Saleiro, Benedict Kuester, Loren Hinkson, Jesse London, Abby Stevens, Ari Anisfeld, Kit T. Rodolfa, and Rayid Ghani. Aequitas: A Bias and Fairness Audit Toolkit. *arXiv 1811.05577*, 2019. URL http://arxiv.org/abs/1811.05577.

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

[154] Kailash Karthik Saravanakumar. The impossibility theorem of machine fairness–a causal perspective. *arXiv preprint arXiv:2007.06024*, 2020.

[155] Robert E Schapire. Explaining adaboost. In *Empirical inference*, pages 37–52. Springer, 2013.

[156] Ralf Schwarzer. Modeling health behavior change: How to predict and modify the adoption and maintenance of health behaviors. *Applied psychology*, 57(1): 1–29, 2008.

[157] Ben Shapiro, Amanda Meng, Cody O'Donnell, Charlotte Lou, Edwin Zhao, Bianca Dankwa, and Andrew Hostetler. Re-shape: A method to teach data ethics for data science education. 04 2020. doi: 10.1145/3313831.3376251.

[158] Michael D Slater and June A Flora. Health lifestyles: Audience segmentation analysis for public health interventions. *Health education quarterly*, 18(2):221–233, 1991.

[159] Girardeau A Spann. Disparate impact. *Geo. LJ*, 98:1133, 2009.

[160] Thilo Spinner, Udo Schlegel, Hanna Schäfer, and Mennatallah El-Assady. explainer: A visual analytics framework for interactive and explainable machine learning. *IEEE transactions on visualization and computer graphics*, 26(1): 1064–1074, 2019.

[161] Shashank Srikant and Varun Aggarwal. Introducing data science to school kids. In *Proceedings of the 2017 ACM SIGCSE technical symposium on computer science education*, pages 561–566, 2017.

[162] Stephen V Stehman. Selecting and interpreting measures of thematic classification accuracy. *Remote sensing of Environment*, 62(1):77–89, 1997.

[163] Holger Steinmetz, Michael Knappstein, Icek Ajzen, Peter Schmidt, and Rüdiger Kabst. How effective are behavior change interventions based on the theory of planned behavior? *Zeitschrift für Psychologie*, 2016.

[164] Siri Steinmo, Christopher Fuller, Sheldon P Stone, and Susan Michie. Characterising an implementation intervention in terms of behaviour change techniques and theory: the 'sepsis six'clinical care bundle. *Implementation Science*, 10:1–9, 2015.

[165] Christina Stoiber, Davide Ceneda, Markus Wagner, Victor Schetinger, Theresia Gschwandtner, Marc Streit, Silvia Miksch, and Wolfgang Aigner. Perspectives of visualization onboarding and guidance in va. *Visual Informatics*, 6(1):68–83, 2022.

[166] Christina Stoiber, Markus Wagner, Florian Grassinger, Margit Pohl, Holger Stitz, Marc Streit, Benjamin Potzmann, and Wolfgang Aigner. Visualization onboarding grounded in educational theories. In *Visualization Psychology*, pages 139–164. Springer, 2023.

[167] Scott A Summers. Sphingolipids and insulin resistance: the five ws. *Current opinion in lipidology*, 21(2):128–135, 2010.

[168] J Tejaswini, T Mohana Kavya, R Devi Naga Ramya, P Sai Triveni, and Venkata Rao Maddumala. Accurate loan approval prediction based on machine learning approach. *Journal of Engineering Science*, 11(4):523–532, 2020.

[169] Scott Thiebes, Sebastian Lins, and Ali Sunyaev. Trustworthy artificial intelligence. *Electronic Markets*, 31:447–464, 2021.

[170] Margery Austin Turner. Mortgage lending discrimination: A review of existing evidence. 1999.

[171] Marialena Vagia, Aksel A Transeth, and Sigurd A Fjerdingen. A literature review on the levels of automation during the years. what are the different taxonomies that have been proposed? *Applied ergonomics*, 53:190–202, 2016.

[172] Isabel Valera, Adish Singla, and Manuel Gomez Rodriguez. Enhancing the accuracy and fairness of human decision making. *Advances in Neural Information Processing Systems*, 31, 2018.

[173] Wil Van Der Aalst and Wil van der Aalst. *Data science in action*. Springer, 2016.

[174] Wil MP van der Aalst, Martin Bichler, and Armin Heinzl. Responsible data science, 2017.

[175] Amelec Viloria, Omar Bonerge Pineda Lezama, and Nohora Mercado-Caruzo. Unbalanced data processing using oversampling: machine learning. *Procedia Computer Science*, 175:108–113, 2020.

[176] Paul Voigt and Axel Von dem Bussche. The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, 10(3152676):10–5555, 2017.

[177] Justice Walker, Sayed Mohsin Reza, Omar Badreddin, Amanda Barany, Karen Del, Karen Del Rio, Alex Acquah, Michael Johnson, Alan Barrera, and Justice Walker. Sandbox data science: Culturally relevant k-12 computing. 2:7, 01 2024. doi: 10.1145/3631986.

[178] Emily Wall, Leslie M. Blaha, Lyndsey Franklin, and Alex Endert. Warning, bias may occur: A proposed approach to detecting cognitive bias in interactive visual analytics. In *2017 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 104–115, 2017. doi: 10.1109/VAST.2017.8585669.

[179] Emily Wall, Arpit Narechania, Adam Coscia, Jamal Paden, and Alex Endert. Left, right, and gender: Exploring interaction traces to mitigate human biases. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):966–975, 2021.

[180] Changhan Wang, Anirudh Jain, Danlu Chen, and Jiatao Gu. VizSeq: A visual analysis toolkit for text generation tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, 2019. doi: 10.18653/v1/D19-3043. URL https://www.aclweb.org/anthology/D19-3043.

[181] Qiaosi Wang, Michael Madaio, Shaun Kane, Shivani Kapania, Michael Terry, and Lauren Wilcox. Designing responsible ai: Adaptations of ux practice to meet responsible ai challenges. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–16, 2023.

[182] Zijie J. Wang, Alex Kale, Harsha Nori, Peter Stella, Mark E. Nunnally, Duen Horng Chau, Mihaela Vorvoreanu, Jennifer Wortman Vaughan, and Rich Caruana. Interpretability, Then What? Editing Machine Learning Models to Reflect Human Knowledge and Values. In *KDD*, 2022. doi: 10.1145/3534678.3539074. URL https://doi.org/10.1145/3534678.3539074.

[183] Zijie J. Wang, Chudi Zhong, Rui Xin, Takuya Takagi, Zhi Chen, Duen Horng Chau, Cynthia Rudin, and Margo Seltzer. TimberTrek: Exploring and Curating Trustworthy Decision Trees with Interactive Visualization. In *VIS*, 2022.

[184] Zijie J Wang, David Munechika, Seongmin Lee, and Duen Horng Chau. Supernova: Design strategies and opportunities for interactive visualization in computational notebooks. *arXiv preprint arXiv:2305.03039*, 2023.

[185] Zijie J. Wang, Aishwarya Chakravarthy, David Munechika, and Duen Horng Chau. Wordflow: Social Prompt Engineering for Large Language Models. *arXiv 2401.14447*, 2024. URL http://arxiv.org/abs/2401.14447.

[186] Zijie J. Wang, Chinmay Kulkarni, Lauren Wilcox, Michael Terry, and Michael

Madaio. Farsight: Fostering Responsible AI Awareness During AI Application Prototyping. In *CHI Conference on Human Factors in Computing Systems*, 2024.

[187] Dean J Wantland, Carmen J Portillo, William L Holzemer, Rob Slaughter, and Eva M McGhee. The effectiveness of web-based vs. non-web-based interventions: a meta-analysis of behavioral change outcomes. *Journal of medical Internet research*, 6(4):e40, 2004.

[188] Vincent Warmerdam, Thomas Kober, and Rachael Tatman. Going beyond T-SNE: Exposing whatlies in text embeddings. In *Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS)*, 2020. doi: 10.18653/v1/2020.nlposs-1.8. URL `https://www.aclweb.org/anthology/2020.nlposs-1.8`.

[189] James Wexler, Mahima Pushkarna, Tolga Bolukbasi, Martin Wattenberg, Fernanda Viegas, and Jimbo Wilson. The What-If Tool: Interactive Probing of Machine Learning Models. *TVCG*, 26, 2019. doi: 10.1109/TVCG.2019.2934619. URL `https://ieeexplore.ieee.org/document/8807255/`.

[190] Isaac Wiafe and Keiichi Nakata. Bibliographic analysis of persuasive systems: techniques; methods and domains of application. In *Persuasive technology: Design for health and safety; the 7th international conference on persuasive technology; PERSUASIVE 2012; Linköping; Sweden; June 6-8; Adjunct Proceedings*, number 068, pages 61–64. Linköping University Electronic Press, 2012.

[191] Rüdiger Wirth. Crisp-dm: Towards a standard process model for data mining. In *Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining*, pages 29–39, 2000.

[192] Jo Wood, Alexander Kachkaev, and Jason Dykes. Design exposition with liter-

ate visualization. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):759–768, 2019. doi: 10.1109/TVCG.2018.2864836.

[193] Peter Xenopoulos, Joao Rulff, Luis Gustavo Nonato, Brian Barr, and Claudio Silva. Calibrate: Interactive Analysis of Probabilistic Model Output. *TVCG*, 29, 2023. doi: 10.1109/TVCG.2022.3209489. URL `https://ieeexplore.ieee.org/document/9904444/`.

[194] Yu Yang, Aayush Gupta, Jianwei Feng, Prateek Singhal, Vivek Yadav, Yue Wu, Pradeep Natarajan, Varsha Hedau, and Jungseock Joo. Enhancing fairness in face detection in computer vision systems by demographic bias mitigation. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '22, page 813–822, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450392471. doi: 10.1145/3514094.3534153. URL `https://doi.org/10.1145/3514094.3534153`.

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

[196] Ashley Zhang, Yan Chen, and Steve Oney. VizProg: Identifying Misunderstandings By Visualizing Students' Coding Progress. In *CHI*, 2023. doi: 10.1145/3544548.3581516.

[197] B. Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 2018. URL `https://api.semanticscholar.org/CorpusID:9424845`.

[198] Tian Zhang, Raghu Ramakrishnan, and Miron Livny. Birch: an efficient data

clustering method for very large databases. *ACM sigmod record*, 25(2):103–114, 1996.

[199] Jianlong Zhou, Amir H Gandomi, Fang Chen, and Andreas Holzinger. Evaluating the quality of machine learning explanations: A survey on methods and metrics. *Electronics*, 10(5):593, 2021.