Behavior Matters: An Alternative Perspective on Promoting Responsible Data Science

ZIWEI DONG, Emory University, USA
AMEYA PATIL, University of Washington, USA
YUICHI SHODA, University of Washington, USA
LEILANI BATTLE, University of Washington, USA
EMILY WALL, Emory University, USA

Data science pipelines inform and influence many daily decisions, from what we buy to who we work for and even where we live. When designed incorrectly, these pipelines can easily propagate social inequity and harm. Traditional solutions are technical in nature; e.g., mitigating biased algorithms. In this vision paper, we introduce a novel lens for promoting responsible data science using theories of behavior change that emphasize not only technical solutions but also the behavioral responsibility of practitioners. By integrating behavior change theories from cognitive psychology with data science workflow knowledge and ethics guidelines, we present a new perspective on responsible data science. We demonstrate example data science interventions in machine learning and visual data analysis, contextualized in behavior change theories that could be implemented to interrupt and redirect potentially suboptimal or negligent practices while reinforcing ethically conscious behaviors. We conclude with a call to action to our community to explore this new research area of behavior change interventions for responsible data science.

CCS Concepts: • **Human-centered computing → HCI theory, concepts and models**; • **Computing methodologies → Machine learning**.

Additional Key Words and Phrases: persuasive technologies, behavior change, human-in-the-loop machine learning, responsible data science, AI ethics

ACM Reference Format:
Ziwei Dong, Ameya Patil, Yuichi Shoda, Leilani Battle, and Emily Wall. 2025. Behavior Matters: An Alternative Perspective on Promoting Responsible Data Science. *Proc. ACM Hum.-Comput. Interact.* 9, 2, Article CSCW034 (April 2025), 23 pages. https://doi.org/10.1145/3710932

1 Introduction

While data science can advance important societal goals, such as fighting climate change and species extinction, it can also cause considerable societal harm [6]. Individual mispredictions can lead to the dehumanization of Black people by labeling them as gorillas [53], or loss of health benefits for those who need them the most [25]. These examples hint towards larger systems of inequity that data science pipelines inadvertently perpetuate when left unchecked.

We have seen a heartening surge in academic research to counteract these inequities in machine learning and broader data science practices [24, 43, 46, 59], including the introduction of the

Authors' Contact Information: Ziwei Dong, ziwei.dong@emory.edu, Emory University, Atlanta, Georgia, USA; Ameya Patil, ameyap2@cs.washington.edu, University of Washington, Seattle, Washington, USA; Yuichi Shoda, yshoda@uw.edu, University of Washington, Seattle, Washington, USA; Leilani Battle, leibatt@cs.washington.edu, University of Washington, Seattle, Washington, USA; Emily Wall, emily.wall@emory.edu, Emory University, Atlanta, Georgia, USA.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

© 2025 Copyright held by the owner/author(s).
ACM 2573-0142/2025/4-ARTCSCW034
https://doi.org/10.1145/3710932

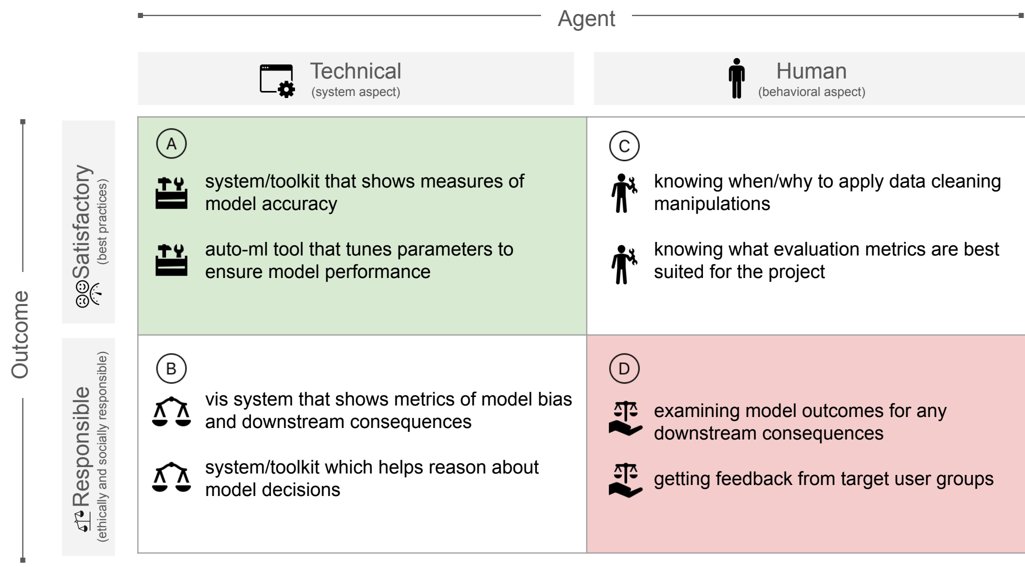

Fig. 1. We characterize data science practices according to desired outcomes (rows – satisfactory and responsible) and agents (columns – technical and human). It is important to note that outcomes are not mutually exclusive. Rigorous data science has historically emphasized technical aspects like auto-tuning 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 paper, 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).

Conference on Fairness, Accountability, and Transparency in 2018, and numerous workshops such as Community-Driven AI: Empowering People Through Responsible Data-Driven Decision-Making at CSCW 2023. Existing work often focuses on ensuring that the data science pipelines, and consequently their outputs, are mathematically and statistically sound. Issues of bias and inequity are then framed as mitigating 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 [12, 78].

However, modifying the algorithms and models that data scientists use is not enough to solve such a systemic problem. We 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 we observe that technical solutions are essential to successful data science we 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 we focus on the *inputs* (data, algorithms) and *outputs* (inferences) and not on the *agents* involved (people and systems), we may miss the opportunity to more meaningfully address the underlying causes of the problems we seek to fix.

The CSCW community has addressed the responsibility of the *agents* involved through exploring and understanding human factors in responsible machine learning and model interpretability [34], the influence of human interaction on the efficacy of ML model-driven decision-making [32], and dissonance in the perception of human and machine understanding [81]. Collaborative data science for social good is of particular interest to the CSCW community [32, 34, 49, 57, 67, 74, 80, 81]. Our work complements these existing efforts by focusing on the human behaviors and interactions that influence responsible data science in both individual and collaborative data science settings. First, data science projects have a broad impact on society and communities[49, 67]. Encouraging responsible data science practices can promote social good at large. Second, data science itself is a community where practitioners must collaborate to deliver responsible models [57, 80]. Ensuring that data science practitioners follow the same responsible behavioral principles fosters collaboration and facilitates the effective delivery of data science projects [74]. By integrating behavior change theories with data science practices, we offer novel methods to support ethical decision-making and collaborative efforts in data science, ultimately contributing to the design, development, and analysis of computer-supported collaborative systems.

In this vision paper, we explore opportunities to formally 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. To ground our discussion, we draw parallels from successful behavior change interventions from cognitive and clinical psychology such as smoking cessation [9] to common data science scenarios today such as model training and exploratory visual data analysis (see Figure 2). In this way, we show how several key principles from foundational behavior change research translate to the data science domain.

Although there have been extensive studies on behavior change approaches [58] and data science ethics [3] separately, we recognize that relatively few projects make *direct* contributions at this intersection of behavior change interventions and data science. Our work culminates in a framework that can help researchers and designers navigate the vast space of prior work relevant to behavior change interventions and apply it in data science contexts. Thus we assert that our work serves as a foundational **call-to-action**: to inspire a critical research agenda focused on cultivating responsible users beyond traditional technical education and training contexts; specifically, through everyday interactions with data science tools. Our objective is to foster reflection among data scientists regarding the significance of their actions toward practicing responsible data science. Finally, we connect our vision to the larger effort to make data science more equitable and just, and outline open challenges for the community moving forward. To summarize, we make the following contributions:

- We introduce the concept of *behavior change interventions for data science*, where we focus on data science behaviors as possible predictors of biased outcomes.

Table 1. A summary of the acronyms used throughout this paper.

Acronym	Meaning
FBC	Factors affecting Behavior Change
BCT	Behavior Change Techniques
MoA	Mechanisms of Action
FBM	Fogg Behavior Model [29]
COM-B	Capability (C), Opportunity (O), and Motivation (M) - Behavior (B) [52]
TDF	Theoretical Domains Framework [4, 50]
BCTT	Behavior Change Techniques Taxonomy [2, 51]

- In section 2, we illustrate how existing psychological models can be applied to inform the design of behavior change interventions in data science. We further introduce how to operationalize them in section 4.
- In section 3, we synthesize a definition of responsible practices in data science work for humans and systems.
- In section 5, we present concrete examples of possible interventions to encourage responsible practices within the data science context.
- We conclude with a discussion of open challenges in the space of behavior change interventions for responsible data science in section 6.

For reference, we include Table 1 below to summarize acronyms that will be introduced and used throughout the remainder of the paper.

1.1 Example Behavior Change Contexts

To ground our forthcoming discussion on behavior change theories, we introduce three concrete examples, shown in Figure 2. Each example is described according to the context of the domain and the agent and desired outcome (as shown in Figure 1). The figure also details the behavior change theories relevant to these contexts and example behavior change interventions, which we will introduce in Sections 2 and 5, respectively.

The first domain (green column) represents a context in which behavior change interventions have been previously employed. In clinical psychology, smoking cessation is a well-known problem that has been approached from numerous perspectives, including that of behavior change [9, 56]. In this case, researchers and clinicians are interested in how to help people live a healthier life by giving up smoking.

The last two domains (blue columns) represent our vision for how we can design behavior change interventions in data science contexts using existing behavior change theories. We cover two example tasks. First, we consider a machine learning context, wherein organizations encourage their data scientists to build accurate, generalizable, and unbiased models. Finally, we consider a visual data analysis context, wherein stakeholders want to ensure that visual data analysis practices and resulting communications are fair and accurate. We use these three example contexts as a running example throughout the paper to review existing behavior change theories, and also to apply the framework provided by these theories for designing interventions for responsible data science.

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 [9, 56] and in environmental domains such as for managing carbon footprint [54, 60]. In spite of the application of behavior change interventions in numerous domains, a survey by Wiafe et al. [76] 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. [55] 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 [14, 50], 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 [29] and Michie et al. [52]. 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

	Clinical Psychology Smoking Cessation	Data Science Psychology Machine Learning	Data Science Psychology Visual Data Analysis
Domain Context (Section 1)	How can we help people who are addicted to smoking, to give up smoking, and lead a healthier life?	How can organizations dealing with data help their data scientists build accurate, truly generalizable and unbiased models?	How can organizations dealing with data help their data scientists create non-misleading visualizations and perform unbiased visual data analysis?
Behavior Change Theories Factor affecting Behavior Change Behavior Change Technique Mechanism of Action (Section 2)	FBM: Trigger COM-B: Opportunity TDF: Environmental context and resources BCTTv1: Antecedent - Restructuring the environment IF: Environmental restructuring Environmental context	FBM: Motivation COM-B: Motivation TDF: Decision Processes BCTTv1: Shaping Knowledge IF: Education/Training Attitudes towards behavior	FBM: Trigger COM-B: Opportunity TDF: Social influences or norms BCTTv1: Feedback and Monitoring IF: Modeling Beliefs about consequences
Agents and Desired Outcomes Technically Satisfactory Behaviorally Responsible (Section 3)	Use cigarettes which reduce the risk of lung cancer Increase self-awareness about health consequences of smoking	Use multiple model performance evaluation metrics Examine model decisions empirically for any downstream consequences	Choose appropriate visual encodings to avoid misinterpretation Evaluate visualizations for dissemination to check if they convey the appropriate message
Interventions (Section 5)	Technically Satisfactory Replace normal cigarettes with those that reduce the risk of lung cancer	Behaviorally Responsible Educate data scientists about the potential bad impacts of model decisions on the target groups	Behaviorally Responsible Prompt visualization designers to evaluate their visualizations for efficacy in conveying the message

Fig. 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 1 and Section 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.

(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 mechanisms that make a specific factor or technique work to influence behavior.

That being said, many theories in each category have overlapping constructs [2, 4, 14, 27, 50, 51] which have been shown to make it difficult to identify individual processes or factors underlying successful behavior change [55]. 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 paper, we instead focus on identifying and discussing the theories that appear most relevant for the data science context, as advised by Pinder et al. [58] and Michie et al. [50]. In this section, we describe these theories and use them to characterize behavior change in the domain contexts listed in Figure 2 to ground them. We demonstrate how to use these theories to generate a series of interventions in section 5.

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** [29] or FBM 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** [52], 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** [4, 50] or TDF identifies 14 empirically verified domains that contain different factors which affect behavior change. TDF [4] 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 [58]. The conciseness and usefulness of the COM-B model is corroborated by the fact that most prior theories [4, 14, 27, 50] including TDF, ultimately break down into components of the COM-B model.

Finally, while COM-B balances specificity and generalizability, the Fogg Behavior Model [29] 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 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.

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.

2.2 Behavior Change Techniques (BCT)

Behavior change techniques put the aforementioned factors of behavior change to work, i.e., implementing the interventions which can bring about behavior change. Michie et al. [52] 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. [2, 51], which lists 93 such techniques clustered into 16 categories. We use the BCTTv1 taxonomy for its descriptive power and Michie et al.'s [52] categorization for its conciseness when designing interventions, as described in our illustrative examples in section 5.

In Figure 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.

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. [15] identified 26 different mechanisms of action and linked the behavior change techniques from the BCTTv1 taxonomy to these mechanisms (e.g., 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 5) with a theoretical grounding, thereby maximizing impact.

In Figure 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 (e.g., 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 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. [18] 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 [17, 47, 66, 75]. In subsection 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 3.2, our primary focus of this paper, elaborated in subsection 3.3, is on the aspect of behavioral responsibility.

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 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 explicit regard to ethics; e.g., 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, e.g., 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 1, we highlight the complementary importance of "Technically Satisfactory" and "Behaviorally Responsible" practices 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. We describe these perspectives in more detail in Sections 3.2 and 3.3 and connect them to the examples in Figure 2.

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 [47].

(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 (e.g., one-way or two-way ANOVA) and parameters like significance level must be carefully considered [7, 37].

(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 [23, 40]. 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 [82]. 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 [19] calls for communicating the results of data analysis sessions with consideration for the data context and uncertainties, especially when using the medium of visualization , 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 [20].

Row 2 in Figure 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 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 [3], Heise et al.'s primary ethical norms in computational research [33], and Zegura et al.'s [79] who called for taking calls for care and social good when practicing data science. Among these literature, we emphasize insights from *Human-Centered Data Science* [3] and the concept of *Care and the practice of data science for social good* [79]. 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* [3] provides a foundational resource that facilitates a systematic approach to contemplating behaviorally responsible practices within the realm of data science. The book offers a set of ethical guidelines that encourage a nuanced consideration of data science projects, e.g., ethics in 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" [3]. 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* [79] 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 [3]. 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** [3, 61, 79]", 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 [49, 79]. For example, Zegura et al. [79] proposed an orientation to a caring mindset in the practice of data science that facilitates social good; Meng et al. [49] 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 [8]. 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 [3, 61, 79], 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 [69]. 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 [70]. 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 [30], 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 [10, 69]. 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) [38] is also a crucial step. Wall et al. [72, 73] 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 [13, 26, 77], 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 [21, 36, 45]. 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 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 in the aforementioned list).

4 Operationalizing Behavior Change Theories for Responsible Data Science

In the preceding sections, we described the 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 science but also decomposing the design process of behavior change interventions in the context of the 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 [4, 29, 50, 52], behavior change techniques [2, 51], and most recently work towards understanding mechanisms of action [15]). In this section, we alter this order to align with how we think about operationalizing these theories toward 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 3.2 and subsection 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 2.1, we need to identify various factors (e.g., *capability, opportunity, and motivation* [52]) 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 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 2), if data scientists lack *motivation* to commit more time to test model outcomes on different influenced groups, this could be bridged by the mechanisms related to changing their *attitudes towards their behaviors* and updating their *beliefs about consequences* [15]. This not only helps designers choose the most appropriate interventions for the digital context but also facilitates maximizing 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 [2, 51] introduced in subsection 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 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 [51]) 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) [65] and specific interventions to their underlying mechanisms of action (MoAs) [16]. To help designers choose appropriate FBCs, MoAs, and interventions, we provide a supplemental table that merges these mappings.

5 Interventions

In this section, we refer back to the two data science contexts in Figure 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 5.1), and visual data analysis example in the third column (subsection 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 4. In subsection 5.3, we provide our own **internal reflection** on the usage of this framework for envisioning behavior change interventions for responsible data science.

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 [68]. 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, which influences 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) [15]** 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) [51]**. 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.a), serving as a constant reminder of the real-world impacts of their work.

Additionally, she **identifies *goals* (MoA) [15]** as another possible underlying mechanism and decides to employ a ***goal priming* (BCT) [51]** 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.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

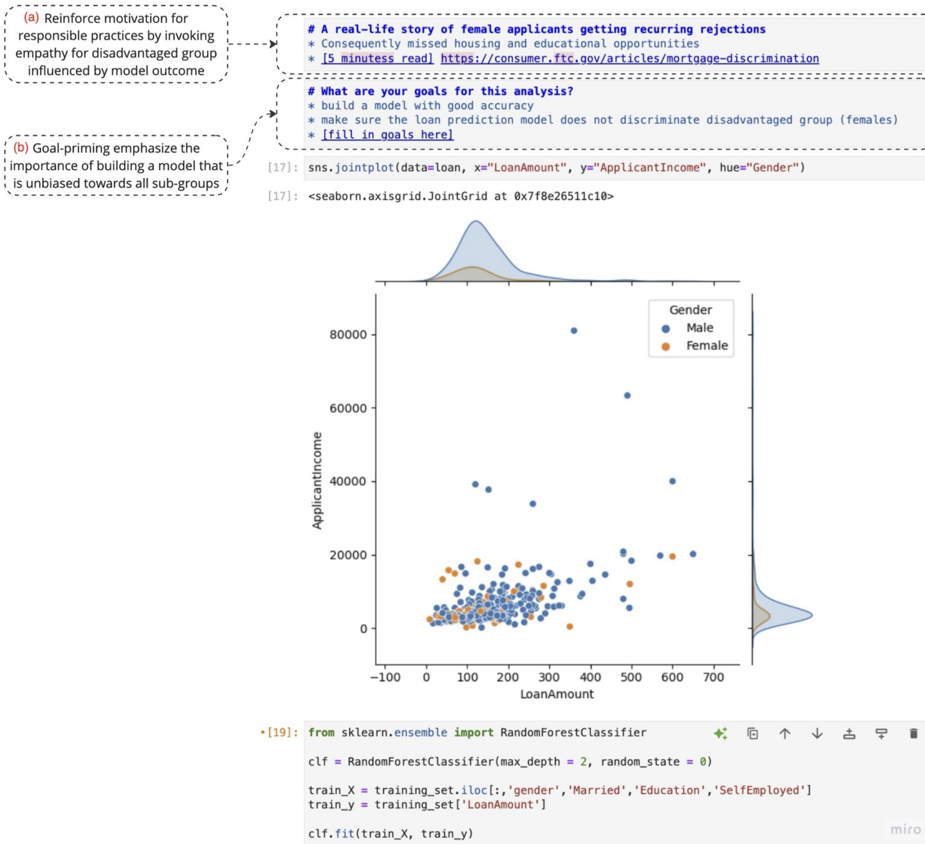

Fig. 3. As data scientists start analyzing the loan approval dataset within a Jupyter notebook, this intervention (a) reinforces their motivation to practice responsible data science by sharing a real-life story that highlights the potential harm that model outcomes can inflict on disadvantaged groups, 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.

envisioning potential interventions using our framework, Maggie ensures that her team is not only technically proficient but also behaviorally responsible.

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 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 maximize impact, Dylan decides to design an intervention tool based on our proposed framework to help the visualization team.

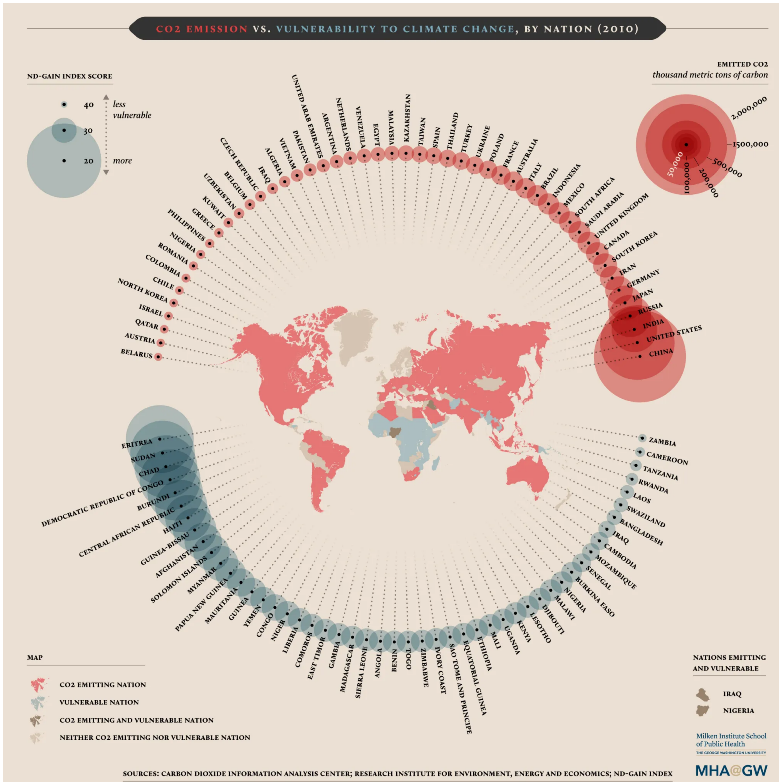

Fig. 4. A data visualization showing which nations are major CO2 emitters, and which nations are vulnerable to the effects of these emissions. In its current state, this visualization might only help global policymakers like the Intergovernmental Panel on Climate Change (IPCC). By gathering feedback 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/

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 **lack of motivation**, and decides to use the *social influences* MoA [51] to help them realize the potential impact on society. Dylan designs interventions to prompt designers to estimate the anticipated range of their target audience [42] (e.g., policymakers, urban public, university students, rural public, etc.). He generates possible scenarios to expose designers to diverse questions audience members might ask about the visualization; e.g., 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)* [51]**. Using our framework, Dylan figures out that gathering feedback from different target audience groups, which falls under ***feedback of behavior*** **(BCT) [51]** 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.

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, 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, e.g., 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 range of behavior change solutions, while leaving the responsibility to choose the right alternative to the practitioner.

6 Open Research Challenges

In this paper, 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.

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 [29], 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, e.g., 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, e.g., 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 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.

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

In this paper, 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 [58]. 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, e.g., provide learning materials like enrolling in a course to *learn new skills*, checklist generation (*planning* and *repetition*), *setting goals* and *tracking progress* [41].

Alternative theories and applications of behavior change interventions emphasize settings intended to encourage longer-term habit formation [28, 64]. Accordingly, we observe at least three major challenges in establishing lasting behavior change 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 [35]. 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 [31, 48], which may be less responsive to intervention. Thus recognizing the boundaries of where interventions may be effective is important for designing more ambitious interventions.

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 [31]? 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.

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 paper (e.g., in subsection 1.1) 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 [51] 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?

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 [39]. 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 [22]. 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 integrate with highly automated data science tools.

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

Throughout this paper, we highlight the importance of education in promoting responsible data science practices. Echoed by many prior works in this space [5, 11, 63], 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 innovative teaching methods, such as experiential learning [71], case studies [44], and interactive simulations [62], 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.

7 Conclusion

In this paper, we introduce the concept of behavior change interventions for data science, emphasizing that data science behaviors can be predictors of biased outcomes. Our work synthesizes a definition of responsible behaviors in data science, encompassing both human (behavioral) and system (technical) aspects. To characterize interventions within data science contexts, we illustrate how existing psychological models can inform the design of behavior change interventions. We operationalize these theories through a four-step framework to help design effective behavior change interventions, which includes (1) identifying problematic and target behaviors, (2) identifying factors affecting problematic behaviors, (3) understanding and employing appropriate Mechanisms of Action, and (4) envisioning potential interventions using Behavior Change Techniques. To inspire the design of practical interventions, we present concrete examples that encourage socially responsible behaviors within the data science context. We conclude by describing the open challenges uncovered by this vision paper and call on our community to explore this emerging research area of behavior change interventions for responsible data science, thereby promoting ethical and socially responsible practices in the field.

Acknowledgement

This work was partially supported by NSF award IIS-2340539

References

[1] 2012. The TRIFACTA data engineering cloud. https://www.trifacta.com/
[2] Charles Abraham and Susan Michie. 2008. A taxonomy of behavior change techniques used in interventions. Health Psychology 27, 3 (2008), 379–387. doi:10.1037/0278-6133.27.3.379

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

[4] 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. 2017. A guide to using the Theoretical Domains Framework of behaviour change to investigate implementation problems. Implementation Science 12, 1 (Dec. 2017), 77. doi:10.1186/s13012-017-0605-9

[5] Teresa K Attwood, Sarah Blackford, Michelle D Brazas, Angela Davies, and Maria Victoria Schneider. 2019. A global perspective on evolving bioinformatics and data science training needs. Briefings in Bioinformatics 20, 2 (2019), 398–404.

[6] Solon Barocas and Danah Boyd. 2017. Engaging the ethics of data science in practice. Commun. ACM 60 (2017), 23 – 25.

[7] Carsten Binnig, Lorenzo De Stefani, Tim Kraska, Eli Upfal, Emanuel Zgraggen, and Zheguang Zhao. 2017. Toward Sustainable Insights, or Why Polygamy is Bad for You. In CIDR.

[8] Ashley Boone, Carl Disalvo, and Christopher A Le Dantec. 2023. 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. 1–13.

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

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

[11] Robert J Brunner and Edward J Kim. 2016. Teaching data science. Procedia Computer Science 80 (2016), 1947–1956.

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

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

[14] James Cane, Denise Oâ€™Connor, and Susan Michie. 2012. Validation of the theoretical domains framework for use in behaviour change and implementation research. Implementation Science 7, 1 (Dec. 2012), 37. doi:10.1186/1748-5908-7-37

[15] Rachel N Carey, Lauren E Connell, Marie Johnston, Alexander J Rothman, Marijn de Bruin, Michael P Kelly, and Susan Michie. 2018. Behavior Change Techniques and Their Mechanisms of Action: A Synthesis of Links Described in Published Intervention Literature. Annals of Behavioral Medicine (Oct. 2018). doi:10.1093/abm/kay078

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

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

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

[19] Michael Correll. 2019. Ethical Dimensions of Visualization Research. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (Glasgow, Scotland Uk) (CHI '19). Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3290605.3300418

[20] Michael Correll and Michael Gleicher. 2014. Error Bars Considered Harmful: Exploring Alternate Encodings for Mean and Error. IEEE Transactions on Visualization and Computer Graphics 20, 12 (Dec. 2014), 2142–2151. doi:10.1109/TVCG.2014.2346298

[21] Henriette Cramer, Jean Garcia-Gathright, Sravana Reddy, Aaron Springer, and Romain Takeo Bouyer. 2019. Translation, Tracks & Data: An Algorithmic Bias Effort in Practice. In Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems (Glasgow, Scotland Uk) (CHI EA '19). Association for Computing Machinery, New York, NY, USA, 1–8. doi:10.1145/3290607.3299057

[22] Anamaria Crisan and Brittany Fiore-Gartland. 2021. 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 (Yokohama, Japan) (CHI '21). Association for Computing Machinery, New York, NY, USA, Article 601, 15 pages. doi:10.1145/3411764.3445775

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

[24] Rachel Eardley, Emma L Tonkin, Ewan Soubutts, Amid Ayobi, Gregory JL Tourte, Rachael Gooberman-Hill, Ian Craddock, and Aisling Ann O'Kane. 2023. Explanation before Adoption: Supporting Informed Consent for Complex Machine Learning and IoT Health Platforms. Proceedings of the ACM on Human-Computer Interaction 7, CSCW1

(2023), 1–25.

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

[26] Mi Feng, Cheng Deng, Evan M. Peck, and Lane Harrison. 2017. HindSight: Encouraging Exploration through Direct Encoding of Personal Interaction History. IEEE Transactions on Visualization and Computer Graphics 23, 1 (2017), 351–360. doi:10.1109/TVCG.2016.2599058

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

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

[29] BJ Fogg. 2009. A Behavior Model for Persuasive Design. In Proceedings of the 4th International Conference on Persuasive Technology (Claremont, California, USA) (Persuasive '09). Association for Computing Machinery, New York, NY, USA, Article 40, 7 pages. doi:10.1145/1541948.1541999

[30] Lisa Gitelman. 2013. "Raw data" is an oxymoron. The MIT Press.

[31] Usman Gohar, Zeyu Tang, Jialu Wang, Kun Zhang, Peter Spirtes, Yang Liu, and Lu Cheng. 2024. Long-Term Fairness Inquiries and Pursuits in Machine Learning: A Survey of Notions, Methods, and Challenges.

[32] Ben Green and Yiling Chen. 2019. The principles and limits of algorithm-in-the-loop decision making. Proceedings of the ACM on Human-Computer Interaction 3, CSCW (2019), 1–24.

[33] 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. 2019. Internet research: Ethical guidelines 3.0. (2019).

[34] Sungsoo Ray Hong, Jessica Hullman, and Enrico Bertini. 2020. Human factors in model interpretability: Industry practices, challenges, and needs. Proceedings of the ACM on Human-Computer Interaction 4, CSCW1 (2020), 1–26.

[35] Eric Horvitz. 1999. Principles of Mixed-Initiative User Interfaces. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (Pittsburgh, Pennsylvania, USA) (CHI '99). Association for Computing Machinery, New York, NY, USA, 159–166. doi:10.1145/302979.303030

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

[37] Eunice Jun, Maureen Daum, Jared Roesch, Sarah Chasins, Emery Berger, Rene Just, and Katharina Reinecke. 2019. Tea: A High-level Language and Runtime System for Automating Statistical Analysis. 591–603. doi:10.1145/3332165.3347940

[38] Daniel Kahneman. 2011. Thinking Fast and Slow. Macmillan Publishers.

[39] Shubhra Kanti Karmaker ("Santu"), Md. Mahadi Hassan, Micah J. Smith, Lei Xu, Chengxiang Zhai, and Kalyan Veeramachaneni. 2021. AutoML to Date and Beyond: Challenges and Opportunities. ACM Comput. Surv. 54, 8, Article 175 (oct 2021), 36 pages. doi:10.1145/3470918

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

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

[42] Elsie Lee-Robbins and Eytan Adar. 2022. Affective Learning Objectives for Communicative Visualizations. IEEE Transactions on Visualization and Computer Graphics (2022), 1–11. doi:10.1109/TVCG.2022.3209500

[43] Cindy Kaiying Lin and Steven J Jackson. 2023. From Bias to Repair: Error as a Site of Collaboration and Negotiation in Applied Data Science Work. Proceedings of the ACM on Human-Computer Interaction 7, CSCW1 (2023), 1–32.

[44] Sabrigiriraj M and K. Manoharan. 2024. Teaching Machine Learning and Deep Learning Introduction: An Innovative Tutorial-Based Practical Approach. WSEAS TRANSACTIONS ON ADVANCES in ENGINEERING EDUCATION 21 (06 2024), 54–61. doi:10.37394/232010.2024.21.8

[45] Michael A. Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. 2020. 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 (Honolulu, HI, USA) (CHI '20). Association for Computing Machinery, New York, NY, USA, 1–14. doi:10.1145/3313831.3376445

[46] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A Survey on Bias and Fairness in Machine Learning. ACM Comput. Surv. 54, 6, Article 115 (jul 2021), 35 pages. doi:10.1145/3457607

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

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

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

[50] S Michie. 2005. Making psychological theory useful for implementing evidence based practice: a consensus approach. Quality and Safety in Health Care 14, 1 (Feb. 2005), 26–33. doi:10.1136/qshc.2004.011155
[51] Susan Michie, Michelle Richardson, Marie Johnston, Charles Abraham, Jill Francis, Wendy Hardeman, Martin P. Eccles, James Cane, and Caroline E. Wood. 2013. 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 (Aug. 2013), 81–95. doi:10.1007/s12160-013-9486-6
[52] Susan Michie, Maartje M Van Stralen, and Robert West. 2011. The behaviour change wheel: a new method for characterising and designing behaviour change interventions. Implementation science 6, 1 (2011), 1–12.
[53] Molly Mulshine. 2015. A major flaw in google's algorithm allegedly tagged two black people's faces with the word'gorillas'. Business Insider (2015).
[54] Kristian S. Nielsen, Sander van der Linden, and Paul C. Stern. 2020. How Behavioral Interventions Can Reduce the Climate Impact of Energy Use. Joule 4, 8 (2020), 1613–1616. doi:10.1016/j.joule.2020.07.008
[55] Rita Orji and Karyn Moffatt. 2018. Persuasive technology for health and wellness: State-of-the-art and emerging trends. Health informatics journal 24, 1 (2018), 66–91.
[56] Jeni Paay, Jesper Kjeldskov, Mikael B. Skov, Lars Lichon, and Stephan Rasmussen. 2015. Understanding Individual Differences for Tailored Smoking Cessation Apps. In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (Seoul, Republic of Korea) (CHI '15). Association for Computing Machinery, New York, NY, USA, 1699–1708. doi:10.1145/2702123.2702321
[57] Samir Passi and Steven J Jackson. 2018. Trust in data science: Collaboration, translation, and accountability in corporate data science projects. Proceedings of the ACM on Human-Computer Interaction 2, CSCW (2018), 1–28.
[58] Charlie Pinder, Jo Vermeulen, Benjamin R Cowan, and Russell Beale. 2018. Digital behaviour change interventions to break and form habits. ACM Transactions on Computer-Human Interaction (TOCHI) 25, 3 (2018), 1–66.
[59] David Piorkowski, Inge Vejsbjerg, Owen Cornec, Elizabeth M Daly, and Öznur Alkan. 2023. AIMEE: An Exploratory Study of How Rules Support AI Developers to Explain and Edit Models. Proceedings of the ACM on Human-Computer Interaction 7, CSCW2 (2023), 1–25.
[60] Henriette Rau, Susanne Nicolai, and Susanne Stoll-Kleemann. 2022. 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). doi:10.3389/fpsyg.2022.901927
[61] L Nelson Sanchez-Pinto, Yuan Luo, and Matthew M Churpek. 2018. Big data and data science in critical care. Chest 154, 5 (2018), 1239–1248.
[62] Ben Shapiro, Amanda Meng, Cody O'Donnell, Charlotte Lou, Edwin Zhao, Bianca Dankwa, and Andrew Hostetler. 2020. Re-Shape: A Method to Teach Data Ethics for Data Science Education. doi:10.1145/3313831.3376251
[63] Shashank Srikant and Varun Aggarwal. 2017. Introducing data science to school kids. In Proceedings of the 2017 ACM SIGCSE technical symposium on computer science education. 561–566.
[64] Holger Steinmetz, Michael Knappstein, Icek Ajzen, Peter Schmidt, and Rüdiger Kabst. 2016. How effective are behavior change interventions based on the theory of planned behavior? Zeitschrift für Psychologie (2016).
[65] Siri Steinmo, Christopher Fuller, Sheldon P Stone, and Susan Michie. 2015. Characterising an implementation intervention in terms of behaviour change techniques and theory: the 'Sepsis Six'clinical care bundle. Implementation Science 10 (2015), 1–9.
[66] Scott Thiebes, Sebastian Lins, and Ali Sunyaev. 2021. Trustworthy artificial intelligence. Electronic Markets 31 (2021), 447–464.
[67] Nenad Tomašev, Julien Cornebise, Frank Hutter, Shakir Mohamed, Angela Picciariello, Bec Connelly, Danielle CM Belgrave, Daphne Ezer, Fanny Cachat van der Haert, Frank Mugisha, et al. 2020. AI for social good: unlocking the opportunity for positive impact. Nature Communications 11, 1 (2020), 2468.
[68] Margery Austin Turner. 1999. Mortgage lending discrimination: A review of existing evidence. (1999).
[69] Wil Van Der Aalst and Wil van der Aalst. 2016. Data science in action. Springer.
[70] Amelec Viloria, Omar Bonerge Pineda Lezama, and Nohora Mercado-Caruzo. 2020. Unbalanced data processing using oversampling: machine learning. Procedia Computer Science 175 (2020), 108–113.
[71] Justice Walker, Sayed Mohsin Reza, Omar Badreddin, Amanda Barany, Karen Del, Karen Del Rio, Alex Acquah, Michael Johnson, Alan Barrera, and Justice Walker. 2024. Sandbox Data Science: Culturally Relevant K-12 Computing. 2 (01 2024), 7. doi:10.1145/3631986
[72] Emily Wall, Leslie M. Blaha, Lyndsey Franklin, and Alex Endert. 2017. 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). 104–115. doi:10.1109/VAST.2017.8585669
[73] Emily Wall, Arpit Narechania, Adam Coscia, Jamal Paden, and Alex Endert. 2021. Left, right, and gender: Exploring interaction traces to mitigate human biases. IEEE Transactions on Visualization and Computer Graphics 28, 1 (2021), 966–975.

[74] Dakuo Wang, Justin D Weisz, Michael Muller, Parikshit Ram, Werner Geyer, Casey Dugan, Yla Tausczik, Horst Samulowitz, and Alexander Gray. 2019. Human-AI collaboration in data science: Exploring data scientists' perceptions of automated AI. Proceedings of the ACM on human-computer interaction 3, CSCW (2019), 1–24.

[75] Qiaosi Wang, Michael Madaio, Shaun Kane, Shivani Kapania, Michael Terry, and Lauren Wilcox. 2023. 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. 1–16.

[76] Isaac Wiafe and Keiichi Nakata. 2012. 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. Linköping University Electronic Press, 61–64.

[77] Jo Wood, Alexander Kachkaev, and Jason Dykes. 2019. Design Exposition with Literate Visualization. IEEE Transactions on Visualization and Computer Graphics 25, 1 (2019), 759–768. doi:10.1109/TVCG.2018.2864836

[78] Yu Yang, Aayush Gupta, Jianwei Feng, Prateek Singhal, Vivek Yadav, Yue Wu, Pradeep Natarajan, Varsha Hedau, and Jungseock Joo. 2022. 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 (Oxford, United Kingdom) (AIES '22). Association for Computing Machinery, New York, NY, USA, 813–822. doi:10.1145/3514094.3534153

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

[80] Amy X Zhang, Michael Muller, and Dakuo Wang. 2020. How do data science workers collaborate? roles, workflows, and tools. Proceedings of the ACM on Human-Computer Interaction 4, CSCW1 (2020), 1–23.

[81] Zijian Zhang, Jaspreet Singh, Ujwal Gadiraju, and Avishek Anand. 2019. Dissonance between human and machine understanding. Proceedings of the ACM on Human-Computer Interaction 3, CSCW (2019), 1–23.

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

Received January 2024; revised July 2024; accepted October 2024