A Design Space of Behavior Change Interventions for Responsible Data Science

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

Teanna Barrett
tb23@cs.washington.edu
University of Washington
Seattle, Washington, 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

ABSTRACT

Behavior change theories, rooted in psychology and sociology, offer valuable insights into *why* and *how* individuals and groups modify their actions and decisions. By leveraging these theories in the context of responsible data science, we can better understand and influence the **behaviors** of data scientists, who play a central role in ensuring ethical outcomes by collecting data, developing, and deploying models. In this paper, we present a comprehensive design space for behavior change interventions aimed at promoting responsible behaviors in data science, structured around the 5W1H interrogative framework (Why, Who, What, When, Where, and How). This framework provides a practical guide for developing effective interventions designed to promote responsible behaviors in data science. We showcase the usability of this design space by using it to characterize existing responsible data science intervention tools. We further demonstrate its utility through two usage scenarios to show how the design space can be applied during the ideation phase for building effective tools to foster responsible data science practices. Our work equips the data science community with resources to create effective interventions that not only ensure technical excellence but also foster ethical responsibility, ultimately benefiting society through the responsible use of data.

CCS CONCEPTS

• **Human-centered computing → Interaction design theory, concepts and paradigms**; *Interaction paradigms.*

KEYWORDS

Behavior Change Intervention, Data Science, HCI, Persuasive Technology

ACM Reference Format:
Ziwei Dong, Teanna Barrett, Ameya Patil, Yuichi Shoda, Leilani Battle, and Emily Wall. 2025. A Design Space of Behavior Change Interventions for Responsible Data Science. In *30th International Conference on Intelligent User Interfaces (IUI '25), March 24–27, 2025, Cagliari, Italy.* ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/3708359.3712140

1 INTRODUCTION

In the evolving landscape of data science, responsible human practices are essential for ensuring ethical data-driven decision-making [69, 72, 73]. The practice of responsible data science is complex, encompassing technical proficiency [33, 62], ethical considerations [6], individual responsibility, and above all the cultivation of a community dedicated to making data-driven decisions that benefit the society at large [70, 84]. At its core, lies the behavior of individuals. For instance, the choices data scientists make in algorithm design can significantly influence the fairness and accuracy of outcomes [69].

Behavior change theories, rooted in psychology and sociology, offer valuable insights into why and how individuals and groups modify their actions and decisions [26]. Leveraging these theories in the context of responsible data science can help us better understand and influence the behaviors of data scientists. Prior work demonstrates the promise of behavior change theories as an avenue for cultivating responsible behaviors in data science [18] . These interventions can effectively nudge, remind, and encourage practitioners to adopt desired responsible practices and reflect on potential biases[27]. Thus, well-designed behavior change interventions have tremendous potential for cultivating a culture of responsibility in data science [37]. However, it is not clear how to operationalize behavior change theories [7, 22, 44, 46] and effective interventions from other fields (e.g., [10, 52, 54, 58]) into interventions for responsible data science.

Towards advancing this vision of responsible data science, **we introduce a design space for behavior change interventions in data science**, illustrated in Figure 1. This design space outlines six critical dimensions that developers can consider when choosing to intervene. The six dimensions are divided into *behavioral considerations* and *implementation considerations* and are based on the 5W1H interrogative framework [29] (**Why**, **Who**, **What**, **When**, **Where**, and **How**). The design space helps developers answer the following questions:

(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?

IUI '25, March 24–27, 2025, Cagliari, Italy

ACM ISBN 979-8-4007-1306-4/25/03
https://doi.org/10.1145/3708359.3712140

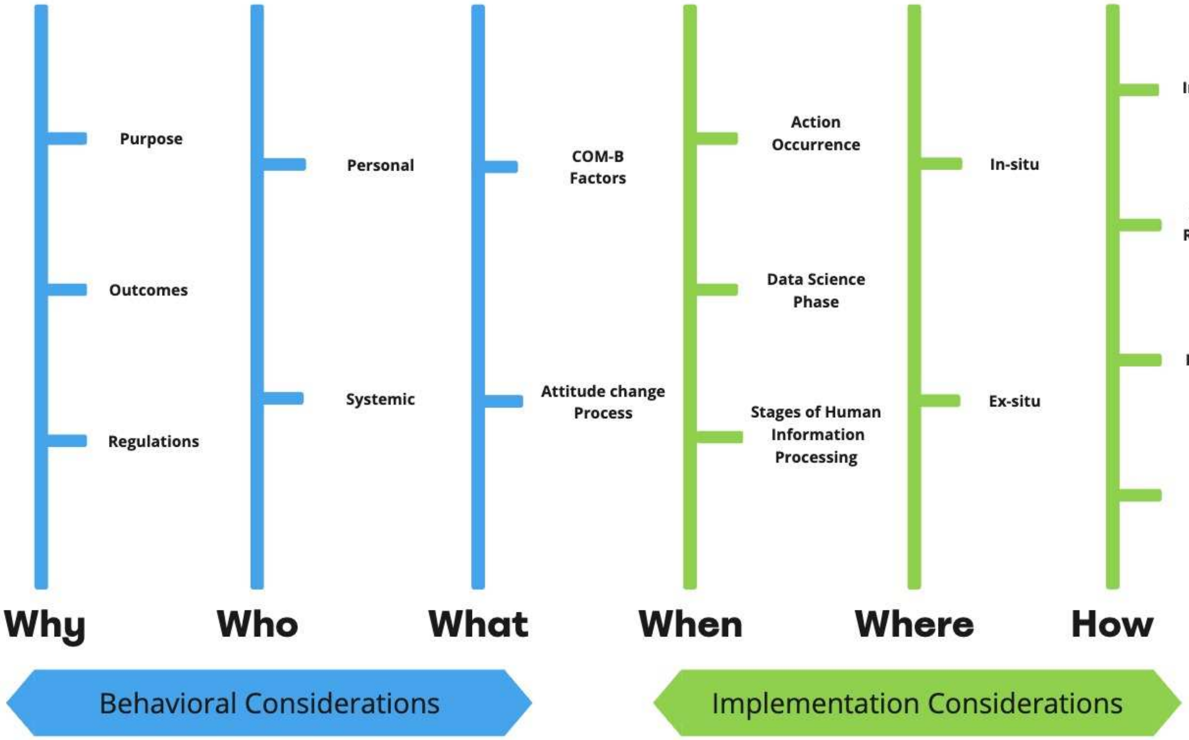


Figure 1: An overview of 5W1H design space proposed in this paper.

(5) **Where** do these interventions take place?
(6) **How** can we design effective interventions?

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 paper 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 [23, 50, 59, 60]; 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*.

2 RELATED WORKS

Dong et al. [18] survey the behavior change literature and highlight three theories relevant to data science, which we summarize as follows: **Factors Affecting Behavior Change (FBC)**, **Behavior Change Techniques (BCT)**, **and Mechanisms of Action (MoA)**.

Factors Affecting Behavior Change (FBC) explains the characteristics that can influence the likelihood of a target behavior being achieved [18]. Established theories introduce key factors that influence behavior change. For example, The Fogg Behavior Model (FBM) identifies three components of behavior: motivation, ability, and trigger. The triggers—spark, facilitator, and signal—play a crucial role in initiating and sustaining behavior change, inspiring targeted strategies and interventions. The COM-B Model emphasizes three factors in behavior change: Capability, Opportunity, and Motivation. Opportunity considers external factors that can facilitate or hinder behavior change.

Behavior Change Techniques (BCT) are targeted strategies or interventions designed to enhance the probability of a desired behavior by utilizing specific influencing factors[18]. BCT applies the previously mentioned factors to create interventions aimed at achieving behavior change. A widely accepted and most detailed taxonomy, Behavior Change Techniques Taxonomy (BCTTv1) [2, 45], enumerates 93 such techniques which are categorized into 16 categories.

Mechanisms of Action (MoA) refers to the cognitive processes that underpin how a particular factor or technique effectively influences behavior[18]. It describes the pathways by which a Behavior Change Technique (BCT) influences behavior, and explains how a behavior change factor impacts a specific technique aimed at achieving the desired change. In the MoA theory introduced by Carey et al., [11], 26 distinct mechanisms of action were identified and were further mapped with relevant behavior change techniques (BCTTv1).

The application of these behavior change theories in responsible data science is a relatively new area, but it draws on a rich body of work from several established fields. In Human-Computer Interaction (HCI), behavior change interventions have been widely explored, particularly in personal health domains, such as smoking cessation [10, 54] and fitness tracking [14]. These interventions target undesirable behaviors and promote positive ones, often using techniques such as nudging, feedback loops, and persuasive technology. Similarly, behavior change models like Fogg's Behavior Model (FBM) [21] and the COM-B Model [46] have been applied to encourage pro-environmental behaviors, such as reducing carbon footprints [52, 58]. In the context of data science, these theories can be applied to develop interventions that encourage ethical data practices and mitigate biases [18]. Recent data science work has only just started to consider this issue, e.g., via notifications of violated fairness and bias metrics [28].

Existing frameworks in responsible data science focus on practitioners by providing checklists or guidelines to follow [23, 50, 59, 60], primarily serving as static resources to guide practitioners on best practices. For example, Rogers et al. [60] focus on providing a checklist for responsible data use in natural language processing, while Saltz et al. [50] offer a systematic framework for ethical considerations in data science projects. However, these frameworks often lack mechanisms for operationalizing ethical principles into actionable interventions that support behavior change.

Furthermore, while other domains have well-established approaches to promoting ethical behavior, it remains unexplored how these theories and other design considerations can be operationalized in interventions for ethical data science practices. This gap emphasizes the need for a structured design space to guide the development of interventions that promote responsible practices within the data science community.

3 DESIGN SPACE RATIONALE

The decision to utilize the 5W1H framework [29] 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 [24, 30, 68] as well as recent applications in Human-Computer Interaction (HCI) and Visualization [32, 65–67]. 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 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 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 4.4 and 5.4. 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.

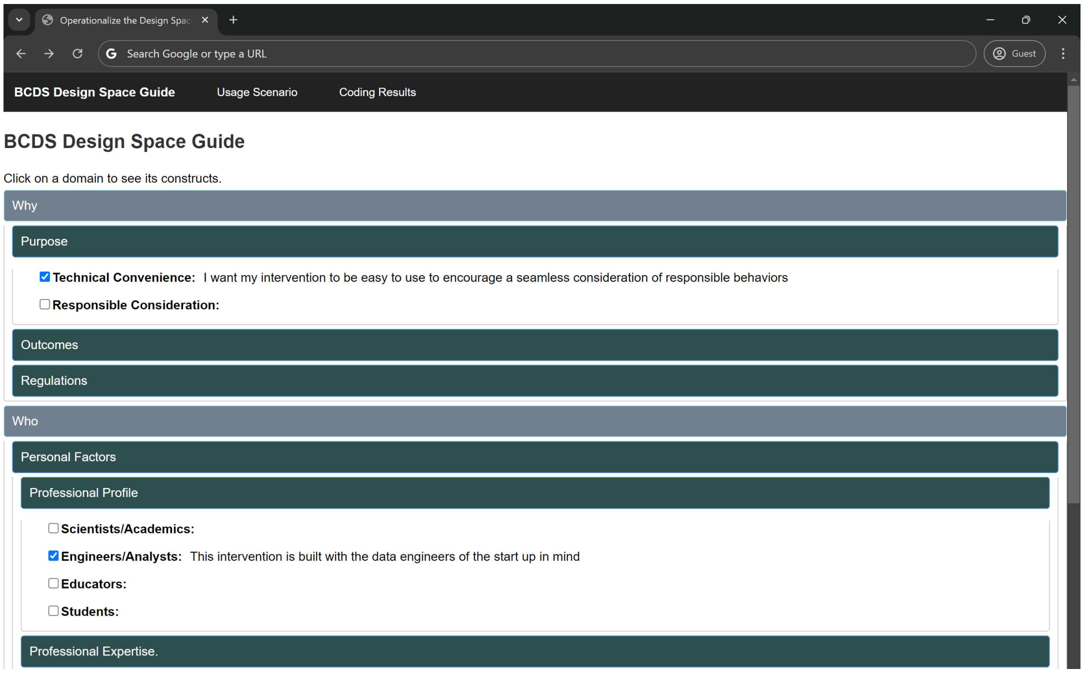
The design space is intended to be a living, collaborative and instructive artifact. Therefore, it can be updated, shared, and referenced for a variety of activities such as documentation, informing stakeholders (especially non-technical), processing design feedback and aiding software developers in 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 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*,



BCDS Design Space Guide

Click on a domain to see its constructs.

Why

Purpose

☑ **Technical Convenience:** I want my intervention to be easy to use to encourage a seamless consideration of responsible behaviors

□ **Responsible Consideration:**

Outcomes

Regulations

Who

Personal Factors

Professional Profile

□ **Scientists/Academics:**

☑ **Engineers/Analysts:** This intervention is built with the data engineers of the start up in mind

□ **Educators:**

□ **Students:**

Professional Expertise.

Figure 2: A screenshot of the interactive BCDS design space website.

- *Who is the target of the behavior change intervention*, and
- *What key objectives does the intervention seek to influence.*

Furthermore, we demonstrate in subsection 4.4 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.

4.1 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 [12]. 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**.

4.1.1 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 [17]). 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.

4.1.2 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.

4.1.3 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 [74] for data privacy in the European Union or HIPAA [3] 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.

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

Interventions that influence behavior must be personalized to their audience to be effective [64]. 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 [72]; 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 [63]. 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.

4.2.1 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 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.

4.2.2 Systemic Factors. Systemic Factors reflect the broader context in which professionals operate, including the organizational and cultural norms that influence their work [5]. 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 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 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 because shifting attitudes can lead to more sustainable and internalized behavior change [42]. To consider what behaviors we aim to change, we orient this dimension with two fundamental questions: **What behavioral factors (COM-B) [46] are being addressed?** and **What attitude change processes [35] are being addressed?**

4.3.1 What Behavioral Factors Are Being Addressed? The COM-B model [46] offers a framework for understanding Behavior (B) as a function of three factors: Capability (C), Opportunity (O), and Motivation (M). The factors that influence a user's behavior the most vary across different scenarios. Identifying which 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) [22]. An intervention might include motivational reminders or gamified elements that reward users for consistent application of ethical practices, thereby boosting motivation.

4.3.2 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 [35]: compliance, identification, and internalization 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 [35]. Compliance typically drives short-term behavior change, as data scientists may comply with data privacy regulations, such as GDPR [57], 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 [35]. 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 [35]. 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.

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

4.4.1 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.

4.4.2 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.

4.4.3 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 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.

5 IMPLEMENTATION CONSIDERATIONS

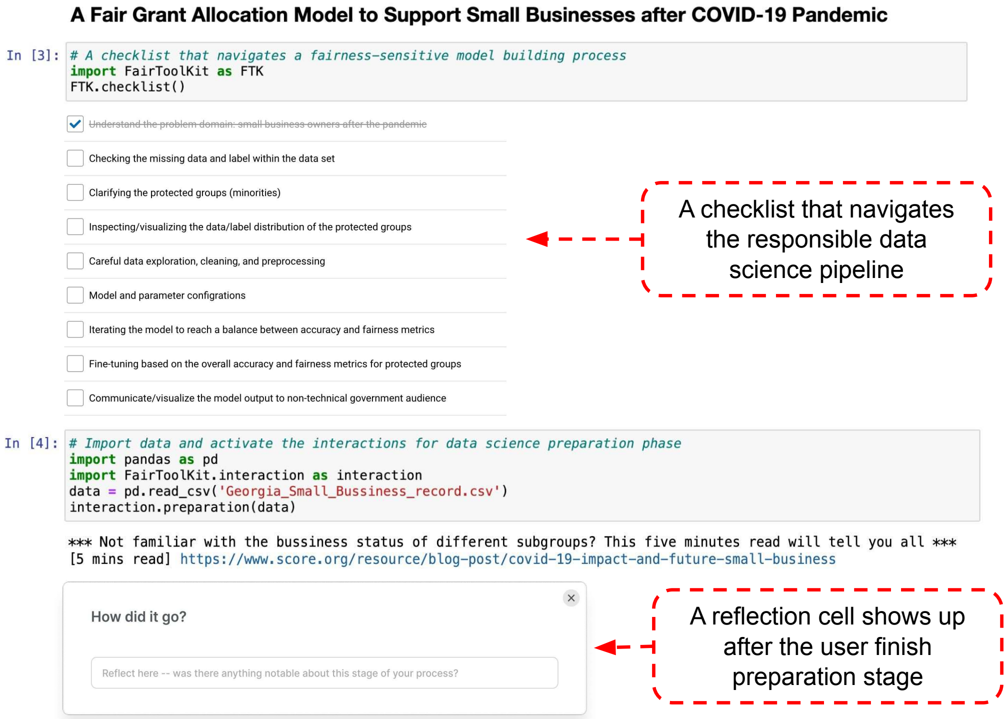
In addition to behavioral factors, the practical and technical logistics of interventions are critical for their success. This section delves into these implementation considerations for the design of behavior change interventions:

- *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 5.4

5.1 When: When is the suitable time to intervene?

The timing of behavior change interventions is a pivotal factor in their effectiveness[13]. 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 [49]. The timing of interventions can be informed by prior work in HCI on intervention and notification timing by Fogarty et al. [20]. Intervention developers can consider at least three different ways of characterizing “When” the intervention occurs: according to the **action occurrence**, the **phase in the**



A Fair Grant Allocation Model to Support Small Businesses after COVID-19 Pandemic

```
In [3]: # A checklist that navigates a fairness-sensitive model building process
        import FairToolKit as FTK
        FTK.checklist()
```

☑ ~~Understand the problem domain: small business owners after the pandemic~~

□ Checking the missing data and label within the data set

□ Clarifying the protected groups (minorities)

□ Inspecting/visualizing the data/label distribution of the protected groups

□ Careful data exploration, cleaning, and preprocessing

□ Model and parameter configrations

□ Iterating the model to reach a balance between accuracy and fairness metrics

□ Fine-tuning based on the overall accuracy and fairness metrics for protected groups

□ Communicate/visualize the model output to non-technical government audience

A checklist that navigates the responsible data science pipeline

```
In [4]: # Import data and activate the interactions for data science preparation phase
        import pandas as pd
        import FairToolKit.interaction as interaction
        data = pd.read_csv('Georgia_Small_Bussiness_record.csv')
        interaction.preparation(data)
```

```
*** Not familiar with the bussiness status of different subgroups? This five minutes read will tell you all ***
[5 mins read] https://www.score.org/resource/blog-post/covid-19-impact-and-future-small-business
```

How did it go?

Reflect here -- was there anything notable about this stage of your process?

A reflection cell shows up after the user finish preparation stage

Figure 3: An exemplary intervention Sean envisioned.

data science process, and the **stage of human information processing**.

5.1.1 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.

5.1.2 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 [15] 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.

5.1.3 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 [55]: (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 [55]. 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 [8]. This stage also includes cognitive operations such as rehearsal, integration, and inference, but these operations occur prior to the point of the decision [55]. 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 [55]. 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 [55]. 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 implement routine fairness checks before any analysis is finalized, ensuring that all outputs adhere to certain ethical standards.

5.2 Where: Where do the interventions take place?

The setting of a behavior change intervention influences its effectiveness [19]. The "Where" dimension analyses how seamlessly interventions integrate into the daily routines of data scientists, influencing their usability and 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 approaches for embedding 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 [36], 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.

5.3 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.

5.3.1 *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.

5.3.2 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).

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

5.3.4 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. [71] and adapt it into the context of data science, as shown in Figure 4. These levels range from complete user control to full automation by the system.

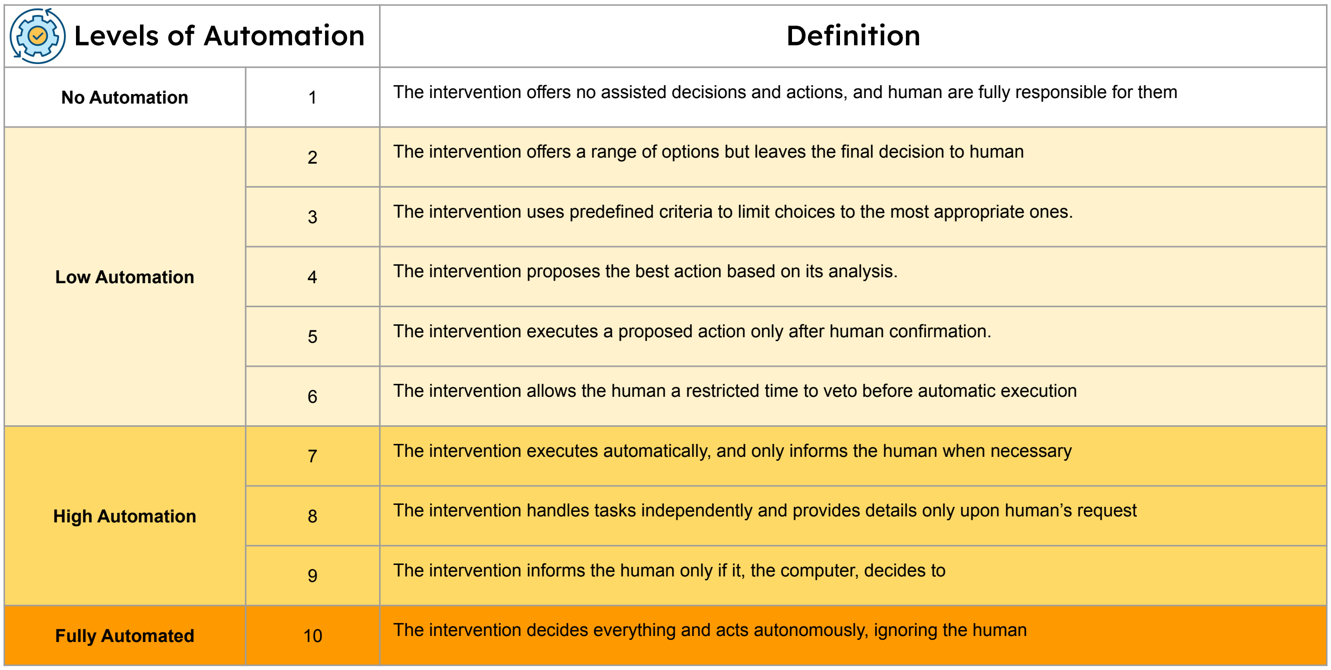
For the sake of simplicity in our subsequent coding of existing responsible data science tools (Section 6), we group the total 10 levels of automation into four types as shown in the Figure 4: No Automation (level 1 in subsubsection 5.3.4), Low Automation(level 2-6 in subsubsection 5.3.4), High Automation (level 7-9 in subsubsection 5.3.4), and Fully Automated (level 10 in subsubsection 5.3.4).

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

5.4.1 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 [1]. 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.

5.4.2 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”).

5.4.3 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. 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 5, the intervention first reminds students to inspect different evaluation metrics with a pop-up box. Furthermore, students



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: We adopt the concept of levels of automation from Vagia et al. [71] to measure the intervention's automation level in the context of data science.

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 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.

6 CHARACTERIZING EXISTING INTERVENTION TOOLS

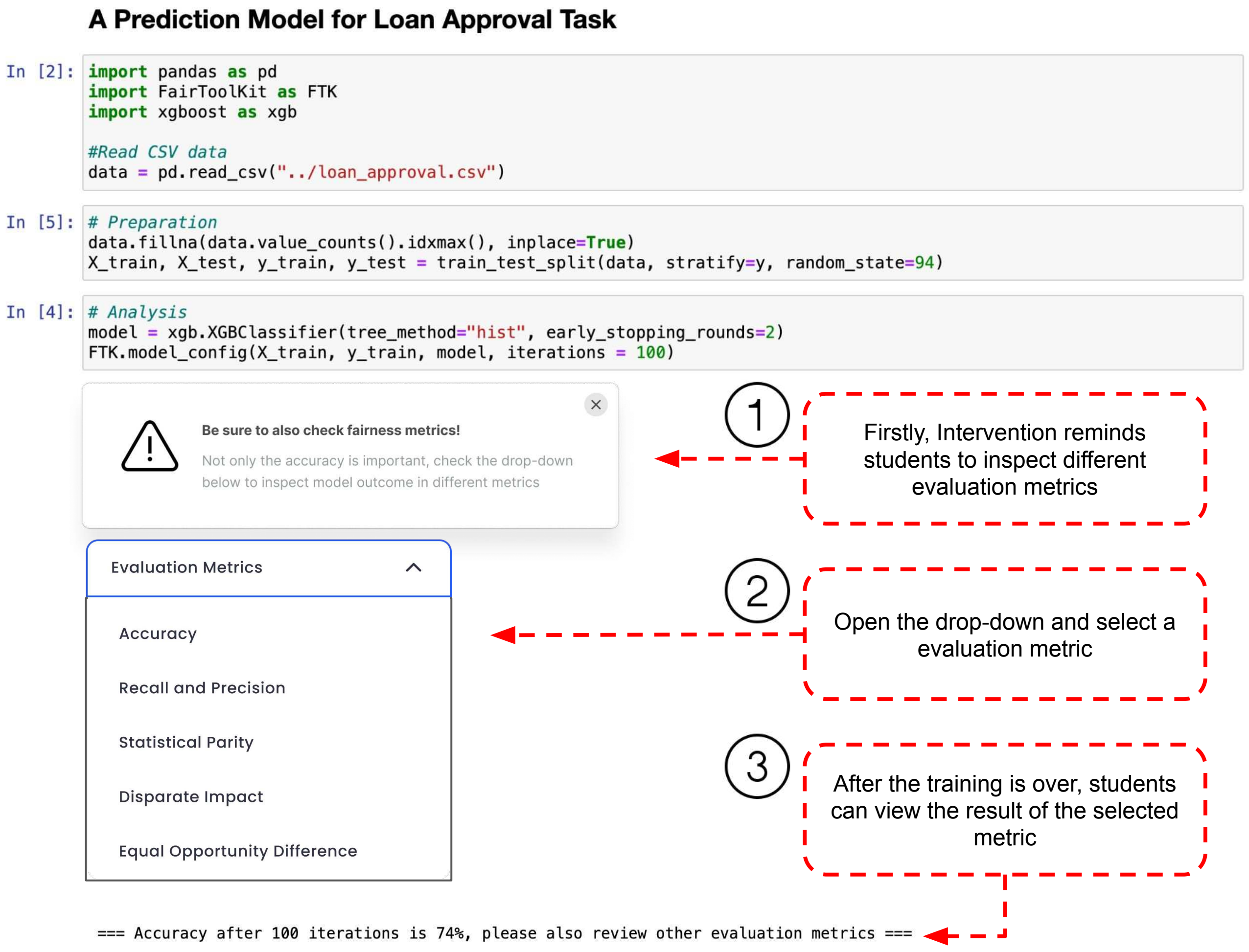
To demonstrate the utility and applicability of our proposed design space, we conducted a targeted survey and coding of existing tools in the domain of RDS. The objective of this analysis is twofold: first, to map the features of these tools to the implementation considerations (When, Where, and How) 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.

6.1 Method

To identify relevant behavior change intervention tools for RDS, we began by reviewing the survey conducted by Wang et al. [79] which covers 163 existing tools that facilitate data science practices (Figure 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 [82]) or ethical considerations (e.g., DocML [9]) 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 [83], bias auditing and reduction [51, 61, 82], model outcome evaluation and monitoring [4, 34, 48, 75], fair model building [40, 77], effective communication through sense-making visualizations [25, 39]; or (2) contributed to ethical considerations more broadly, such as by providing machine learning documentation for ethical priming during model deployment [9, 28, 85], highlighting the consequences of configuration changes on fairness [38, 76, 78], or explaining model interpretability [47, 53, 80, 81].

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



A Prediction Model for Loan Approval Task

```
In [2]: import pandas as pd
        import FairToolKit as FTK
        import xgboost as xgb

        #Read CSV data
        data = pd.read_csv("../loan_approval.csv")
```

```
In [5]: # Preparation
        data.fillna(data.value_counts().idxmax(), inplace=True)
        X_train, X_test, y_train, y_test = train_test_split(data, stratify=y, random_state=94)
```

```
In [4]: # Analysis
        model = xgb.XGBClassifier(tree_method="hist", early_stopping_rounds=2)
        FTK.model_config(X_train, y_train, model, iterations = 100)
```

Be sure to also check fairness metrics!

Not only the accuracy is important, check the drop-down below to inspect model outcome in different metrics

Firstly, Intervention reminds students to inspect different evaluation metrics

Open the drop-down and select a evaluation metric

After the training is over, students can view the result of the selected metric

```
=== Accuracy after 100 iterations is 74%, please also review other evaluation metrics ===
```

Figure 5: An exemplary intervention Dr. Y envisioned.

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 [43] 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.

6.2 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 7 provides a visual representation of these results, with key findings highlighted below:

F1: *(How)* Minimal automation to develop responsible skillset. As described in subsubsection 5.3.4 and Figure 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 the data scientist (e.g., DocML [9] only reminds

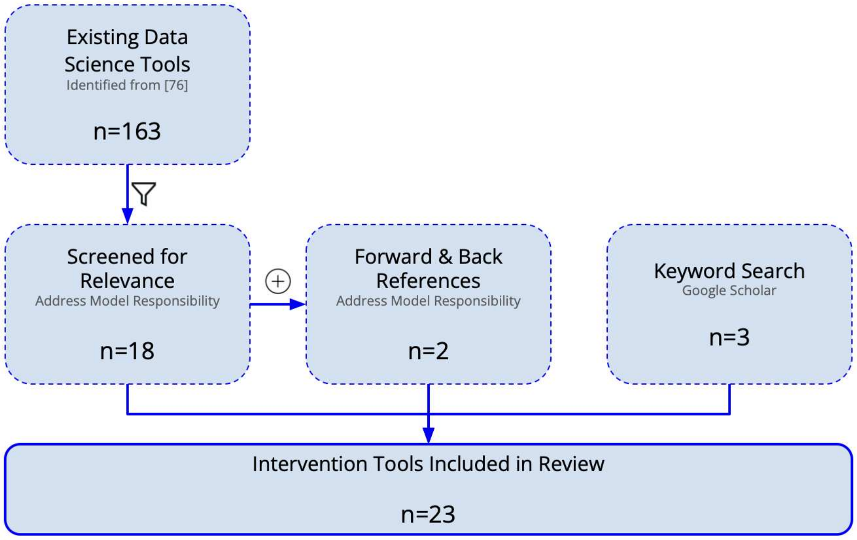

Existing Data Science Tools
Identified from [76]
n=163
Screened for Relevance
Address Model Responsibility
n=18
Forward & Back References
Address Model Responsibility
n=2
Keyword Search
Google Scholar
n=3
Intervention Tools Included in Review
n=23

Figure 6: The process of arriving at the 23 interventions discussed in Section 6.

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 [40] 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 [16].

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.

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. However, this also highlights a potential area for improvement, as critical ethical considerations may sometimes require 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 subsubsection 4.3.2). 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.

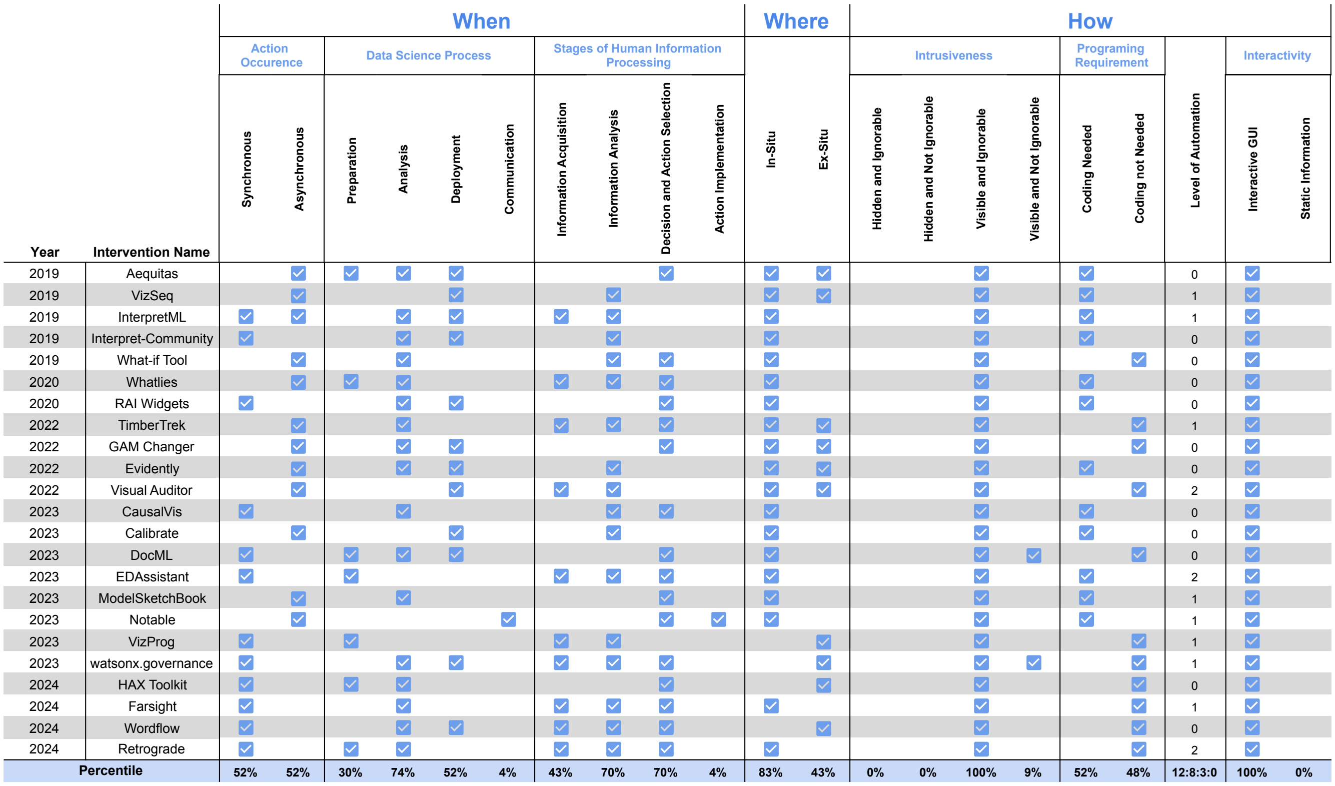
Takeaway: We encourage the development of non-ignorable interventions, especially for high-stakes analysis scenarios.

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[56]. 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.** 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 [41]. 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



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

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 [83].
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 lifecycle. 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 [80] helps users to summarize different levels of the decision tree model at scale) and 4% of interventions support Action Implementation (e.g., Notable [39] 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 [39]. 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 [53]). 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 [31]. 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.

7 DISCUSSION AND CONCLUSION

Generalizability and Robustness: While this framework was validated with a specific 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.

Limitations: While our proposed design space offers a framework for understanding and guiding behavior change interventions in responsible data science, it is not without its limitations. First, the design space is built primarily on existing theories and models, which may not fully capture the rapidly evolving nature of data science practices and technologies. Additionally, our framework focuses on currently available interventions, meaning it may overlook emerging tools or techniques that could present new opportunities or challenges in promoting ethical practices. Another limitation lies in the generalizability of our findings, as our examples and case studies are largely contextualized within specific selected data science tools and environments, potentially limiting their applicability to other domains or broader application contexts. Lastly, while we emphasize the need for both behavioral and implementation considerations, the relative importance of each factor may vary depending on the specific use case or organizational context, which our framework does not explicitly address.

Conclusion: In this paper, we explore 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 paper 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 functionalities 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.

REFERENCES

[1] 2020. South German Credit. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5QG88.

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

[4] Evidently AI. 2022. Evidently: Evaluate and Monitor ML Models from Validation to Production. Evidently AI. https://github.com/evidentlyai/evidently

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

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

[7] 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

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

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

[10] 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.

[11] 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

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

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

[14] Sunny Consolvo, David W McDonald, and James A Landay. 2009. Theory-driven design strategies for technologies that support behavior change in everyday life. In *Proceedings of the SIGCHI conference on human factors in computing systems.* 405–414.

[15] Anamaria Crisan, Brittany Fiore-Gartland, and Melanie Tory. 2021. Passing the Data Baton : A Retrospective Analysis on Data Science Work and Workers. *IEEE*

Transactions on Visualization and Computer Graphics 27, 2 (2021), 1860–1870. doi:10.1109/TVCG.2020.3030340

[16] Wesley Hanwen Deng, Manish Nagireddy, Michelle Seng Ah Lee, Jatinder Singh, Zhiwei Steven Wu, Kenneth Holstein, and Haiyi Zhu. 2022. Exploring How Machine Learning Practitioners (Try To) Use Fairness Toolkits. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 473–484. doi:10.1145/3531146.3533113

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

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

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

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

[21] 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

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

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

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

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

[26] Martin S Hagger, Linda D Cameron, Kyra Hamilton, Nelli Hankonen, and Taru Lintunen. 2020. *The handbook of behavior change*. Cambridge University Press.

[27] Pelle Guldborg Hansen and Andreas Maaløe Jespersen. 2013. Nudge and the manipulation of choice: A framework for the responsible use of the nudge approach to behaviour change in public policy. *European journal of risk regulation* 4, 1 (2013), 3–28.

[28] Galen Harrison, Kevin Bryson, Ahmad Emmanuel Balla Bamba, Luca Dovichi, Aleksander Herrmann Binion, Arthur Borem, and Blase Ur. 2024. 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*. 1–19.

[29] Geoff Hart. 2002. The five w's of online help systems. *Geoff Hart* (2002).

[30] Sylvia C Hewitt. 2006. The five W's of progesterone receptors A and B: now we know where and when. *Endocrinology* 147, 12 (2006), 5501–5502.

[31] Simon David Hirsbrunner, Michael Tebbe, and Claudia Mueller-Birn. 2024. From critical technical practice to reflexive data science. *CONVERGENCE-THE INTERNATIONAL JOURNAL OF RESEARCH INTO NEW MEDIA TECHNOLOGIES* 30, 1 (FEB 2024), 190–215. doi:10.1177/13548565221132243

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

[33] Jessica Hullman and Andrew Gelman. 2021. Designing for interactive exploratory data analysis requires theories of graphical inference. *Harvard Data Science Review* 3, 3 (2021), 10–1162.

[34] IBM. 2020. watsonx.governance. IBM. https://www.ibm.com/products/watsonx-governance

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

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

[37] Gerjo Kok. 2018. A practical guide to effective behavior change: How to apply theory-and evidence-based behavior change methods in an intervention. (2018).

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

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

[40] Xingjun Li, Yizhi Zhang, Justin Leung, Chengnian Sun, and Jian Zhao. 2023. EDAssistant: Supporting Exploratory Data Analysis in Computational Notebooks with In Situ Code Search and Recommendation. *ACM TiiS* 13 (2023). doi:10.1145/3545995

[41] Harry D. Mafukidze, Action Nechibvute, Abid Yahya, Irfan Anjum Badruddin, Sarfaraz Kamangar, and Mohamed Hussien. 2024. Development of a Modularized Undergraduate Data Science and Big Data Curricular Using No-Code Software Development Tools. *IEEE ACCESS* 12 (2024), 100939–100956. doi:10.1109/ACCESS.2024.3429241

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

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

[44] 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

[45] 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

[46] 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.

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

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

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

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

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

[52] 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

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

[54] 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

[55] R. Parasuraman, T.B. Sheridan, and C.D. Wickens. 2000. 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 (2000), 286–297. doi:10.1109/3468.844354

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

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

[58] 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

[59] Shaina Raza, Shardul Ghuge, Chen Ding, Elham Dolatabadi, and Deval Pandya. 2024. FAIR Enough: Develop and Assess a FAIR-Compliant Dataset for Large Language Model Training? *Data Intelligence* 6, 2 (2024), 559–585.

[60] Anna Rogers, Tim Baldwin, and Kobi Leins. 2021. Just What do You Think You're Doing, Dave?'A Checklist for Responsible Data Use in NLP. *arXiv preprint arXiv:2109.06598* (2021).

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

[62] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. 2021. "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. 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 39, 15 pages. doi:10.1145/3411764.3445518

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

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

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

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

[67] Christina Stoiber, Markus Wagner, Florian Grassinger, Margit Pohl, Holger Stitz, Marc Streit, Benjamin Potzmann, and Wolfgang Aigner. 2023. Visualization Onboarding Grounded in Educational Theories. In *Visualization Psychology*. Springer, 139–164.

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

[69] Linnet Taylor and Nadezhda Purtova. 2019. What is responsible and sustainable data science? *Big Data & Society* 6, 2 (2019), 2053951719858114.

[70] 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.

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

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

[73] Wil MP van der Aalst, Martin Bichler, and Armin Heinzl. 2017. Responsible data science. 311–313 pages.

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

[75] Changhan Wang, Anirudh Jain, Danlu Chen, and Jiatao Gu. 2019. 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*. doi:10.18653/v1/D19-3043

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

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

[78] Zijie J. Wang, Chinmay Kulkarni, Lauren Wilcox, Michael Terry, and Michael Madaio. 2024. Farsight: Fostering Responsible AI Awareness During AI Application Prototyping. In *CHI Conference on Human Factors in Computing Systems*.

[79] Zijie J Wang, David Munechika, Seongmin Lee, and Duen Horng Chau. 2023. SuperNOVA: Design Strategies and Opportunities for Interactive Visualization in Computational Notebooks. *arXiv preprint arXiv:2305.03039* (2023).

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

[81] Vincent Warmerdam, Thomas Kober, and Rachael Tatman. 2020. Going beyond T-SNE: Exposing Whatlies in Text Embeddings. In *Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS)*. doi:10.18653/v1/2020.nlposs-1.8

[82] James Wexler, Mahima Pushkarna, Tolga Bolukbasi, Martin Wattenberg, Fernanda Viegas, and Jimbo Wilson. 2019. The What-If Tool: Interactive Probing of Machine Learning Models. *TVCG* 26 (2019). doi:10.1109/TVCG.2019.2934619

[83] Peter Xenopoulos, Joao Rulff, Luis Gustavo Nonato, Brian Barr, and Claudio Silva. 2023. Calibrate: Interactive Analysis of Probabilistic Model Output. *TVCG* 29 (2023). doi:10.1109/TVCG.2022.3209489

[84] 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.

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