Designing Effective Training Dataset Explanations: The Impact of Information Depth and Progressive Disclosure

Ariful Islam Anik
Computer Science Department
University of Manitoba
Winnipeg, Manitoba, Canada
anikmai@myumanitoba.ca

Andrea Bunt
Computer Science Department
University of Manitoba
Winnipeg, Manitoba, Canada
bunt@cs.umanitoba.ca

Abstract

Transparency in AI is crucial for fostering user trust and acceptance, yet achieving it through explanations presents significant design challenges, particularly regarding how much detail to provide. For example, in-depth explanations can convey accurate and comprehensive information, but they also risk overwhelming users. This paper considers this important design tradeoff in the context of training dataset explanations, which describe the data used to train AI systems and differ from most model-centric explanations in terms of what and how much information they communicate. Specifically, we investigate how information depth in training dataset explanations and the use of Progressive Disclosure impact users' understanding of an AI system (assessed via their critiques of the system), their system assessments, and their cognitive load. Findings from a study with 32 participants show advantages to providing users with comprehensive information on training datasets. Detailed explanations not only enhanced perceived trust, fairness, and understanding, but were also preferred by participants despite the increased cognitive load. While Progressive Disclosure did not effectively mitigate cognitive load, it improved users' perception of learning. These findings suggest that effective transparency does not come from minimizing detail, but from embracing it, as participants consistently valued clarity and completeness over brevity, even at the cost of higher cognitive load.

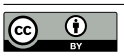
CCS Concepts

• **Human-centered computing** → HCI design and evaluation methods; User studies.

Keywords

Transparency, Explanations, Training Datasets, Information Depth, Progressive Disclosure

ACM Reference Format:
Ariful Islam Anik and Andrea Bunt. 2026. Designing Effective Training Dataset Explanations: The Impact of Information Depth and Progressive Disclosure. In *31st International Conference on Intelligent User Interfaces (IUI '26), March 23–26, 2026, Paphos, Cyprus*. ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3742413.3789087



IUI '26, Paphos, Cyprus

ACM ISBN 979-8-4007-1984-4/2026/03
https://doi.org/10.1145/3742413.3789087

1 Introduction

Transparency is essential for the responsible use of Artificial Intelligence (AI) systems across different domains (e.g., healthcare [30, 50], criminal justice [4, 19, 20], hiring [69, 82, 104]), as it enables users to understand and evaluate how AI systems make decisions, which in turn promotes trust, accountability, and effective human–AI collaboration [18, 48, 64, 70, 103, 127, 133]. Yet achieving transparency is challenging because AI systems often function as black boxes [67, 77, 98]. This opacity has driven the field of Explainable AI (XAI) to develop techniques that make algorithmic decision-making more interpretable and accessible to users [1, 3, 18, 39, 48, 56, 103, 111].

Despite major progress, a predominant focus in XAI research has been on the technical aspects of model transparency [1, 16, 42, 44, 75, 76, 101], which can overlook other dimensions that are equally important to understanding how AI systems operate [41, 88]. The Human-Computer Interaction (HCI) community has therefore argued for a broader and more human-centric approach to XAI (i.e., HCXAI) that prioritizes users' goals, needs, and understanding [2, 41, 44, 47, 75]. For example, one critical dimension is training data, as datasets used to train AI systems fundamentally shape their behavior, biases, and limitations [10, 86]. Providing transparency on training data [16, 60, 134] can provide critical context to help users predict and interpret a system's behavior. This recognition has led to the development of training dataset explanations [5, 6], which draw on dataset documentation [53] to describe data sources, collection methods, demographics, and intended use. While studies show that such explanations promote transparency [5], they also raise important design questions about how to present them effectively to users [6, 16].

Within the broader human-centered perspective of HCXAI, the appropriate level of detail to provide within an explanation is a key design dilemma. Concise explanations are easier to process and often preferred by receivers [78, 88, 105], but they risk oversimplifying complex systems leading to incomplete understanding [111]. In contrast, detailed explanations can enhance users' comprehension and contribute to more accurate impressions of the system [32, 68], but at the cost of greater cognitive load [68]. In other words, the level of detail has meaningful impacts on core aspects of human-AI collaboration (e.g., system impressions and understanding), yet the inconclusive findings leave the appropriate amount of detail an open question. For training dataset explanations, this question remains especially unclear for two primary reasons. First, these explanations tend to be much longer than model-centric explanations (e.g., LIME [106], SHAP [81]) because of the richness of data-related information they convey. Second, their intended use case differs, as they are typically presented during onboarding [6, 24, 25], when

users are first introduced to a system rather than during repeated interactions. This unique context makes it important to understand the tradeoffs between the information depth and cognitive load to identify what level of detail best supports meaningful and comprehensible transparency.

In this paper, we investigate how different levels of information depth (i.e., the level of detail and complexity included within an explanation) impact the utility and impact of training dataset explanations. We explore this through a user study where we ask participants to critique two automated systems (i.e., assess their potential advantages and disadvantages [6]) using training dataset explanations that varied in information depth (summary vs. detailed). We examine how information depth impacts users' perceptions, understanding, cognitive load, and system critiques. We also examine whether Progressive Disclosure, a design technique that reveals explanation components on-demand, while keeping track of what has already been explained [115], can help manage the complexity of explanations across different depths.

Our study with 32 participants revealed that in-depth training dataset explanations resulted in higher perceived trust, fairness, understanding of the data, learning, and cognitive load in comparison to summarized explanations. Progressive Disclosure, however, had a statistically significant effect only on participants' perception of learning but showed no significant reduction in cognitive load. Regarding participants' critiques, both explanations supported productive evaluation of the systems, but the focus of their critiques varied depending on information depth. Our interviews revealed a clear preference for in-depth explanations despite the additional cognitive load, indicating that participants valued understanding the system clearly even with the increased effort.

Our work provides the following contributions. We provide empirical evidence on how information depth in training dataset explanations shapes users' comprehension, system assessments, and cognitive load. We also highlight trade-offs between information depth and cognitive load during onboarding and offers preliminary insights into how Progressive Disclosure can enhance the perception of learning. Together, these contributions advance the broader agenda of human-centered XAI by highlighting design considerations for designing explanations that are both detailed and accessible in onboarding contexts.

2 Related Work

2.1 Overview of Different XAI Approaches

The field of XAI has made notable progress in producing a growing collection of explanations of AI systems. Common explanation approaches include *local explanations* that explain individual AI system decisions (e.g. counterfactual [106, 109], case-based explanations [24, 95]) and *global explanations* that help users to understand the overall system (e.g., feature importance [18, 38], feature contribution [9, 32, 128, 133]). However, majority of these explanations are algorithm-centered [41, 42, 75]. Recognizing the limitations of a solely algorithm-centered focus [34, 45, 101], researchers have called for a shift to broader contexts [88, 89]. These calls gave rise to socio-technical approaches, where the focus of explanations includes factors outside of the algorithmic black-box [40, 44, 46, 88], and data-centric approaches, where the focus of the explanations is on the training dataset of the system [5, 13, 16, 17, 60, 84, 134].

Data-centric approaches present properties of the training data to shed light on data that is influencing model predictions and to allow receivers to detect biases and inconsistencies [16, 17]. Training dataset explanations [5] are a data-centric approach that communicate important information about datasets, such as the motivation, creation, composition, intended uses, distribution, and maintenance of a dataset. As they build upon the dataset documentation literature [53], they share common ground with artifacts such as Data Cards [102] and Model cards [36, 90]. However, unlike documentation artifacts, which primarily target dataset creators and direct dataset consumers (e.g., machine learning developers) [53, 84], training dataset explanations are explicitly designed to be understandable by end users [5]. Training dataset explanations are also in line with calls for broader views of explanations that include socio-technical factors beyond the model [5, 41, 53]. Prior work has shown potential for such explanations to improve system transparency in an onboarding scenario [5, 6], by helping users reflect on the system before interacting with it directly [6, 24, 25]. Our work further explores the feasibility of training dataset explanations as an onboarding tool, with particular emphasis on how varying the information depth affects their utility.

2.2 Impact of AI Explanations on Users

Numerous studies have shown that explanations can promote transparency [5, 35, 103], inform people's trust in [64, 77, 111] and acceptance of the systems [35, 66, 131], and influence fairness perceptions [5, 15, 39, 73, 111]. However, the impact of AI explanations is not universally positive. Some studies reported no impact of explanations on trust [32, 35, 101], suggesting gaps between the focus of explanations and users' needs.

In some cases, explanations can even prompt users to act against their own interests, leading to unanticipated and unintended negative consequences. Ehsan and Riedl defined these negative effects as "explainability pitfalls" [45]. Our synthesis of existing literature suggests two primary circumstances under which explanations could inadvertently result in such pitfalls. The first arises from the unintended impact of explanations on users, including uncalibrated trust [45, 63, 75], overreliance [9, 21, 75, 101], high cognitive load [68, 115, 122], misinterpretation [6], and frustration [45, 75, 76]. For example, users may develop uncalibrated trust [63, 75], accepting AI outputs without sufficient critical evaluation if the explanation appears clear or convincing, leading to overreliance [22, 99]. Overly detailed or complex explanations can contribute to high cognitive load [68, 111], overwhelming users and leading to frustration [45], misinterpretation [6], and even overreliance [23]. The second circumstance is directly related to explanation design, including distracting interfaces [76], designs that promote passive consumption of information [76], lack of actionability [45, 75, 76], excessive information volume and complexity [68, 76], and ambiguity [76]. For example, if explanations are not designed with the users' information or visual literacy in mind [8, 112], they can lead to misinterpretation and improper trust in the system. Similarly, a lack of actionability can leave users unsure of how to apply

the information to improve outcomes [41, 74], which can cause frustration [45].

The existence of these negative consequences suggests that further systematic study of a range of explanation types and design parameters is required to build a comprehensive understanding of how AI explanations can better support users. We contribute to this body of work by investigating the design of training dataset explanations, with a particular focus on the role of information depth.

2.3 Information Depth as a Factor of AI Explanation Design

Several factors influence users' interactions with and perceptions of AI explanations, including the type [39, 111], depth [32, 68, 111], modality [122], presentation style of information [6, 15], and user characteristics [43, 122]. Information depth, which refers to the level of detail and complexity included in the content of an explanation, stands out as a critical design factor. Prior work has produced mixed results concerning the impact of information depth on user understanding [32, 68, 111, 115], explanation preference [88], system reliance [23, 99], and cognitive load [31, 64, 68].

These mixed results can be interpreted through Cognitive Load Theory [120], which distinguishes intrinsic, extraneous, and germane load [119, 121]. Increasing explanation depth can raise intrinsic load by introducing more concepts, dependencies, and relationships that users must process [121]. At the same time, when detailed explanations are well-structured and clearly articulated, they may reduce extraneous load by clarifying system behavior and minimizing ambiguity. Depth can also support germane load by encouraging users to reflect on system behavior and construct more coherent mental models. Since intrinsic, extraneous and germane load are considered to be additive [97], reducing extraneous load while increasing germane load is only effective if the overall cognitive demands remain within users' processing limits [116, 121]. When explanation depth exceeds available cognitive resources, increases in intrinsic load may not be offset by reductions in extraneous load or gains in germane processing. In such cases, information depth can overwhelm users, increasing the risk of shallow understanding, misinterpretation, or reliance on cognitive shortcuts [97, 100, 124]. This perspective helps explain why information depth can both improve understanding and increase cognitive effort, and why its effects depend on the amount of information provided and the way it is structured for users.

The above theoretical tension is reflected in empirical findings and recommendations in the literature. For example, some researchers argue that AI explanations should be selective, drawing on properties of human explanation [79], because complete reasoning can overwhelm users [71, 88]. Simple explanations are often preferred by receivers [78, 88, 105], but can lead to incomplete understanding [111]. Other researchers argue for prioritizing comprehensibility of explanations over their compactness (i.e., limiting information to avoid overwhelming users) [76]. In-depth information in explanations can help users to develop the most accurate mental models [68, 111], but can also lead to higher cognitive load [31, 64, 68]. Moreover, detailed explanations can lead users resort to cognitive shortcuts [123], such as accepting explanation material without critical reflection or selectively focusing on details that confirm their pre-existing beliefs (i.e., confirmation bias [65]), contributing to overreliance [23, 99]. Excessive detail can also lead to false beliefs about the system's accuracy [115]. These mixed results indicate a lack of clear guidelines on how much detail to include in explanations.

This gap is especially relevant for Training dataset explanations [5, 6], which are typically longer than other types of AI explanations studied to date. For example, prior work explored four different types of explanations [18, 39] (input-influence, sensitivity, case-based, demographic), all of which contained fewer than 100 words. Schoeffer et al. explored different combinations of common explanation types (relevant factors, factor importance, and counterfactual scenarios) [111], with the most comprehensive explanation consisting of 172 words. In contrast, in prior work on training dataset explanations [6], the shortest explanation was over three times the length at 613 words, reflecting the breadth of information they communicate on dataset properties. This raises open questions regarding how established findings on information depth generalize to long-form training dataset explanations. There remains limited guidance on how much information is beneficial in training dataset explanations, particularly for onboarding contexts where users are forming an initial mental model of the AI system. We contribute to this gap by exploring the impact of two levels of information depth on users' comprehension and usage of training dataset explanations.

2.4 Explanation Presentation Paradigms and Progressive Disclosure

Beyond what information is presented, how explanations are presented to users plays a critical role in shaping their impact [15, 107, 122]. Prior work distinguishes between different explanation presentation paradigms [7], including static explanations, where users are presented with all available explanatory information at once [7, 18, 33], and on-demand, interactive explanations, where users actively choose when and what explanatory content to reveal [16, 33, 67, 129]. Static explanations can support rapid formation of initial mental models [68, 77], but may also increase cognitive load and overreliance [9, 68]. In contrast, on-demand explanations can support intentional engagement and critical reflection [22], while risking underuse by users who lack motivation [122].

One interface mechanism for enabling on-demand access is Progressive Disclosure, a concept from UI design that involves hiding advanced interface controls, limiting initial user errors while they are learning the system [29, 92]. In the context of AI explanations, Progressive Disclosure has been proposed as a mechanism for providing effective transparency by balancing transparency with cognitive manageability [115]. Effective Progressive Disclosure typically requires that (1) information is revealed on demand, (2) explanation content is organized hierarchically, and (3) the system keeps track of the context and information already given to users [115]. In practice, there is limited empirical evidence showing that Progressive Disclosure reliably reduces cognitive load or improves system perceptions. Moreover, presenting explanations on demand may shift cognitive effort from processing content to deciding when

additional information is needed, which may introduce new forms of cognitive cost [93, 100].

In this study, we contrast a static presentation of the full training dataset explanation with an on-demand presentation of content implemented through Progressive Disclosure. This allows us to examine not only how information depth affects users' comprehension and critique, but also how the mode of explanation delivery shapes perceived learning and cognitive effort in training dataset explanations.

3 Study on Information Depth and Progressive Disclosure

We conducted a study to gain insights into how varying the amount of information in training dataset explanations influence participants' subjective impression of the system, their understanding of both the system and the explanation, and their cognitive load. We further investigated whether Progressive Disclosure helps manage participants' cognitive load.

3.1 Participants

We recruited participants through various channels, including advertisements across a university campus, on different online platforms such as X (formerly Twitter) and LinkedIn, and through snowball sampling. To ensure diversity in participants in terms of AI knowledge and background, we used a questionnaire to prescreen participants. The questionnaire included questions on participants' academic background (computer science/ engineering/ non-engineering) and formal AI training (formal courses taken on AI/ML). We also included the AI literacy scale [126], which consists of twelve 7-point Likert items that cover four core constructs of AI literacy: awareness, usage, evaluation, and ethics. We recruited 34 participants but excluded two who did not follow the study procedure or complete the full study. Therefore, our final sample included 32 participants (14 men, 17 women, 1 non-binary). Our sample size is consistent with other human-centered XAI studies where the primary data is qualitative [17, 35, 41, 42, 122]. With qualitative data requiring manual coding, having a manageable data volume is an important consideration [26]. Moreover, human-centered XAI studies involving larger participant pools (e.g., [9, 22, 32, 73, 133]) typically feature shorter participant engagement (e.g., 10-30 minutes), whereas our study sessions averaged two hours per participant. Regarding participants' experience with ML and AI, 24 participants had no formal training (i.e., academic or professional experience) with ML and AI, while 8 did. In terms of academic background, 17 had a CS background, 6 had an Engineering background, and 9 came from non-CS / Engineering fields. The mean AI literacy score for our participant pool was 5.6 (SD = 0.73) out of a maximum of 7. The demographics and background questions can be found in the auxiliary materials.

3.2 Study Design

We included two main factors in our study.

- Information Depth: Summary vs. Detailed
- Progressive Disclosure: Present vs. Absent

Information Depth refers to the level of detail and complexity provided in the explanation. We defined two levels for this factor. In the *summary* version, the explanation offers high-level, concise information about the training data, while in the *detailed* version, the information is presented in a more elaborate and precise manner. Section 3.3 describes the information present in the two levels of this factor. The second factor is the *Progressive Disclosure* of information [115], which is a mechanism that enables transparency "on demand". We also defined two levels for this factor: the *presence* or *absence* of *Progressive Disclosure*.

Our study had a mixed 2 x 2 design with *Information Depth (summary, detailed)* as the within-subjects factor and *Progressive Disclosure (absent, present)* as the between-subjects factor. In other words, each participant interacted with explanations with both levels of *Information Depth* and one level of *Progressive Disclosure*. The order of *Information Depth* was counterbalanced across participants. We chose *Information Depth* as a within-subjects factor to enable participants to directly compare *summary* and *detailed* explanations. We prioritized eliciting contrastive comments for this factor given the trade-offs identified prior work [78, 88, 105]. We were also concerned about mitigating the impact of individual differences (e.g., AI Literacy [126], prior AI training) with this factor. While *Progressive Disclosure* could also be interesting as a within-subject factor, a fully within-subjects design was not practical with the length of the study sessions (averaging two hours).

3.3 Explanation Interfaces and Content

Our training dataset explanation, following prior work [6], presented four categories of information. These categories include: i) *Collection* (the collection process, data sources, and pre-processing of the data); ii) *Demographics* (the demographic distribution of the data instances); iii) *Recommended Usage* (list of recommended use cases); and iv) *General Information* (overview information about the dataset, including release date, history of usage, and updates).

For the *detailed* explanations, we used the same Q&A design and content as prior work [6] (see Figure 1 (A) and Figure 1 (B)). The questions and answers include information on sample size, attributes, data collection process, data pre-processing process, demographic information in multiple dimensions (e.g., age, gender), recommended use, release dates, history of usage, and updates. The *summary* versions contained four concise summaries (one for each category), which presented only a subset of the information with the aim of conveying the core message (see Figure 1 (C) and Figure 1 (D)). For example, regarding the data collection process, the *detailed* explanation included how an automated extraction tool was initially used to gather and pre-process the data, followed by manual evaluation of a subset to verify the accuracy. In contrast, the *summary* explanation simply stated that the data collection process combined automated and manual approaches.

Regarding *Progressive Disclosure*, when it was *present*, we applied the principles of Progressive Disclosure (as described in Springer and Whittaker [115]) within the interface. This was achieved by adding interactivity, such as keeping only one category open at a time and requiring users to click on questions to view the answers, with the viewed answers remaining open. In the *absence* of *Progressive Disclosure*, all information was initially visible, without the

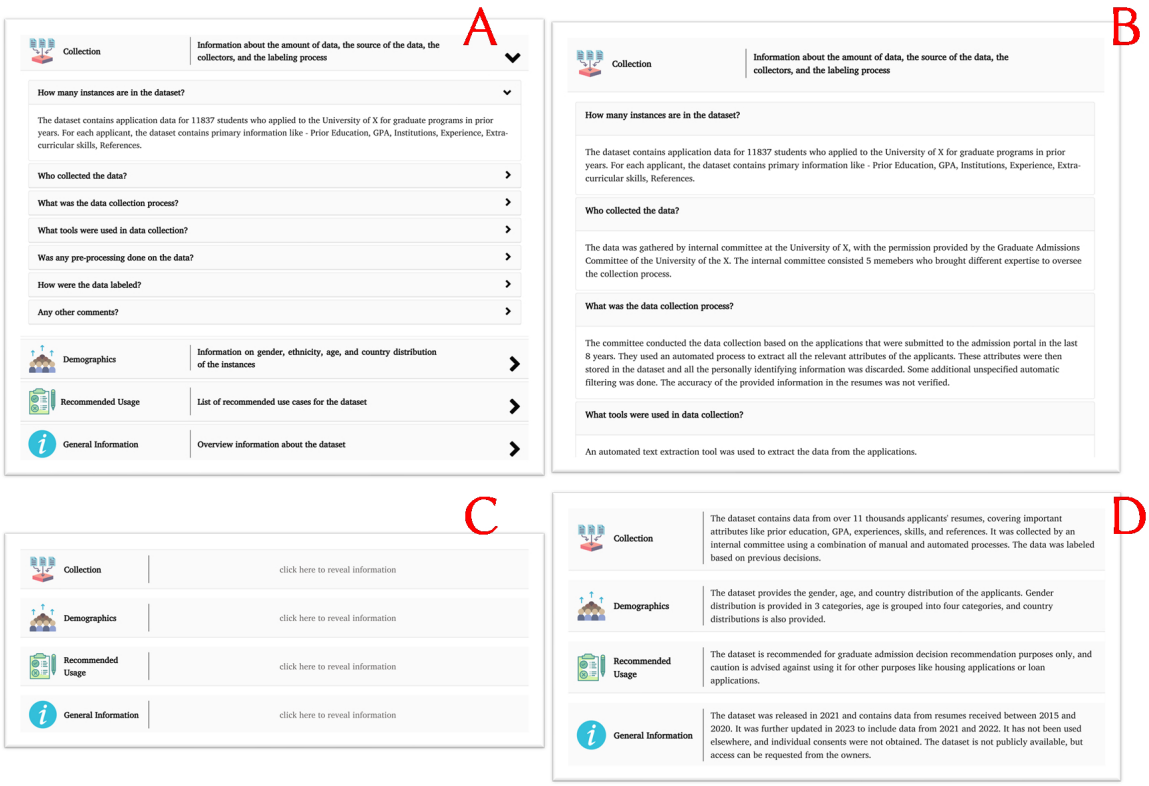


Figure 1: Snippets from the different versions of training dataset explanations. (A) demonstrates a *detailed* explanation with progressive disclosure *present* whereas (B) demonstrates a snippet from a *detailed* explanation with progressive disclosure *absent*. (C) demonstrates a *summary* explanation with progressive disclosure *present* and (D) demonstrates a *summary* explanation with progressive disclosure *absent*. All explanations can be found in full in the auxiliary materials.

additional interactivity required to view content. Our approach is consistent with existing literature on how to achieve progressive disclosure [114, 136], prioritizing a design that allows users to selectively explore information aligned with their interests. Figure 1 (A) and Figure 1 (C) depict the two explanations where *Progressive Disclosure* was *present,* whereas Figure 1 (B) and Figure 1 (D) demonstrate versions where *Progressive Disclosure* was *absent.* Complete versions of the explanations can be found in the auxiliary materials.

3.4 Scenario and Task

We used a scenario-based approach [108] where we asked participants to critique two AI systems based on training dataset explanations with each level of *Information Depth.* We used two scenarios from prior studies on training dataset explanations [5, 6] with the aim of having scenarios of similar significance and familiarity to a diverse audience: an automated hiring system scenario and an automated admission system scenario.

Participants were asked to imagine that they work in the HR department of a company (for the automated hiring scenario) or are part of the recruitment committee at a university (for the admission system scenario) and they were tasked with utilizing the training dataset explanation to recommend whether to purchase the system. As part of this recommendation, participants were asked to use the information in the explanation to critique the training data and the system. To guide participants in the critique, we used a simplified version of the SWOT analysis technique [14, 59] where we asked participants to identify strengths and weaknesses of the data and the system [6]. To encourage in-depth critiques, we asked participants to provide around eight comments with a minimum of one strength and one weakness, however, we did not impose a strict limit and told them to critique as they saw fit. We counterbalanced the scenario order across participants. The detailed scenarios can be found in the auxiliary materials.

3.5 Study Procedure

We conducted our study sessions remotely over Zoom, with one researcher present to conduct the study. Following the introduction of the study, we asked participants to complete a brief demographics questionnaire. Subsequently, we provided them with the task description and demonstrated the tool (Userback) [135] that enabled them to annotate interface elements and enter critique comments. Participants were then presented with one of the scenarios and an explanation interface (either *summary* or *detailed*), which they used to critique the system. Upon completing their first critique,

participants completed a post-scenario questionnaire comprised of Likert-scale questions assessing their trust in the system, perception of fairness, perceived understanding of the explanation, perceived learning about ML, and cognitive load. The post-scenario questionnaire also contained five comprehension questions (designed to gauge participants' understanding) about the data and the system based on information found in the explanations: two multiple-choice questions (e.g., Was there any update provided to the dataset?) and three open-ended questions (e.g., What kind of processes were used in data collection?).

After completing the first scenario, participants proceeded through the same process for the second scenario with an explanation interface with the other level of *Information Depth*. Upon completion of both scenarios, we conducted a semi-structured interview with each participant to gain more detailed insights into their critique and their perceptions of the differing levels of *Information Depth*. The interview included open-ended questions prompting participants to reflect on how they interpreted the explanations, which aspects they found helpful or challenging, how the level of detail influenced their critique, and how they compared the two explanation interfaces. The study sessions lasted about 2 hours on average. Participants spent approximately 70 minutes on the critique tasks, 20 minutes on questionnaires, and 25 minutes on the interview. Participants were compensated $30 for their time. The study was approved by our institutional research ethics board. All study materials can be found in the auxiliary materials.

3.6 Data Collection and Analysis

We collected two primary sources of data. The first was questionnaire responses on how participants felt about the data, the system, and the explanation through a Likert-scale based post-scenario questionnaire. Additionally, we collected data on their cognitive load via the NASA TLX tool [58]. We used Wilcoxon Signed-Rank tests to analyze within-subject comparisons of *Information Depth* and Mann–Whitney U tests to analyze between-subject comparisons by *Progressive Disclosure*. We calculated Cronbach's α for measures with combined items to check for internal consistency, using a threshold of 0.70 to determine acceptable reliability. We used a non-parametric test since we did not assume a normal distribution.

The second primary data source was participants' critique comments. Two researchers were involved in analyzing the critique data. The researcher who conducted the study sessions first coded the data using a coding scheme from a prior study of training dataset explanations [6]. During this phase, the researcher applied the coding scheme to categorize each critique based on the specific information they referred to within the explanation and their accuracy. The second researcher participated in the analysis through regular meetings with a focus on discussing ambiguous critiques and alternative interpretations. When disagreement arose, we revisited the critique alongside the relevant explanation content and discussed our reasoning until we reached agreement. This iterative process continued till we reached agreement on the coding decisions for all critiques. Our process emphasized coding in a way that meaningfully captured the intent of participants' critiques, rather than on maximizing agreement as a metric. Therefore, in line with qualitative analysis guidelines in HCI research [85], we did not calculate an Inter-rater Reliability score for our qualitative data coding process.

When comparing means, we used a repeated measures ANOVA with *Information Depth* as the within-subject factor and *Progressive Disclosure* as the between-subject factor. We ran Pearson's chi-squared test to compare the categorical distributions. We used $p = .05$ as the significance threshold and $p < .1$ as an indication of a trend (i.e., approaching significance [27]).

As a supplement to these primary data sources, we collected qualitative data through semi-structured interviews, which we audio recorded and later transcribed for analysis. Two researchers were involved in the analysis. We conducted a directed qualitative content analysis [61] to examine participants' discussions of *Information Depth* and *Progressive Disclosure*. Following a deductive approach, the researcher who conducted the study sessions coded relevant comments in which participants explicitly talked about these factors. Both researchers then discussed the coded data over multiple meetings to refine the findings. We chose this approach to provide additional qualitative context to our critique and questionnaire data while maintaining a focused exploration of the study variables.

4 Result

4.1 Impact of Information Depth and Progressive Disclosure on User Perceptions

As illustrated in Table 1, *Information Depth* impacted participants' subjective impressions of the utility of the explanation content and their impressions of the system. Participants felt they understood the data more with the *detailed* explanation (Mdn = 17, IQR = 3) than they did with the *summary* explanation (Mdn = 11.5, IQR = 10.75; Z = 4.155, $p < .001$). Figure 2 (left) shows that there was much greater variability for this measure with the *summary* explanation than there was with the *detailed* version. With the *detailed* explanation, participants further perceived the information as more sufficient to critique the system (Z = 3.987, $p < .001$) and felt that they learned more (Z = 3.355, $p < .001$). The *detailed* explanation also impacted participants' perceptions in the system, in that they reported higher levels of trust in the system (Z = 3.091, $p = .002$) and that the system was fairer (Z = 3.913, $p < .001$). Finally, unsurprisingly, participants reported higher perceptions of depth (Z = 3.086, $p = .002$) and cognitive load (Z = 2.872, $p = .004$) with the *detailed* explanation. The fact that the medians are on the low end of the 7-point scale suggests that even the *detailed* explanation was not perceived as too detailed or overwhelming.

In terms of *Progressive Disclosure,* as both conditions provided the same information, expectedly so, there were no differences in the first five measures in Table 1. However, as demonstrated in Table 2, there was a statistically significant impact on participants' perception of *learning* (Z = 2.515, $p = .012$), with participants' perceived learning being higher when *Progressive Disclosure* was *present* (Mdn = 8.75, IQR = 3.25) than when it was *absent* (Mdn = 6.75, IQR = 2.5). Additional tests separating the data for each *Information Depth* revealed a trend for *summary* (Z = 1.651, $p = .099$) and a significant difference for *detailed* (Z = 2.395, $p = .017$) explanation. The increased perception of learning with the *presence*

Table 1: Median (IQR) values for the Likert-scale questionnaire data based on Information Depth. We report the median and IQR values since we did not assume a normal distribution of the Likert-scale data and conducted non-parametric test. We also provide scale ranges as some measures combine multiple questionnaire items.

	Scale Range	Summary	Detailed	Z	Sig
Trust	8-56	31.5 (8.75)	36.5 (7.75)	3.091	.002
Fairness	4-28	13 (6.75)	16.5 (5.75)	3.913	<.001
Perceived understanding of the data	3-21	11.5 (10.75)	17 (3)	4.155	<.001
Perception of depth	1-7	1.5 (2)	3 (2)	3.086	.002
Sufficient information to critique	3-21	9 (6)	15 (6)	3.987	<.001
Perceived learning	2-14	6 (5.75)	9 (3.75)	3.355	<.001
Cognitive load	1-7	2.5 (3)	3.5 (2)	2.872	.004

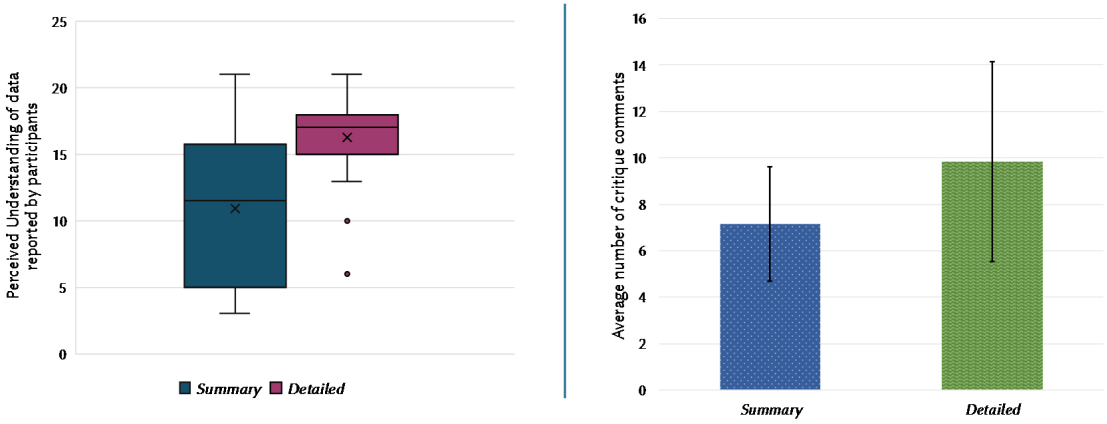


Figure 2: Left - Participants' perceived understanding of the data across Information Depth (the sum of responses to three items); Right - Average number of comments per participant across Information Depth. Error bars show standard deviation.

Table 2: Median (IQR) values for the Likert-scale questionnaire data based on Progressive Disclosure.

	Scale Range	Absent	Present	Z	Sig
Perceived learning	*2-14*	*6.75 (2.5)*	*8.75 (3.25)*	*2.515*	*.012*
Cognitive load	1-7	3 (1.38)	3 (2)	0.609	.543

of *Progressive Disclosure* supports findings from earlier research which showed *Progressive Disclosure* affords more efficient learning progress [29]. For cognitive load, none of the explanations induced a high cognitive load on participants (each with a median of 3), and the use of *Progressive Disclosure* did not statistically significantly impact this measure.

4.2 Impact of Information Depth and Progressive Disclosure on Participants' Critiques

To understand how *Information Depth* and *Progressive Disclosure* impacted the nature of participants' objective assessment of the system, we analyzed their critique comments across three dimensions: the volume of their critique comments, the breadth of their critique topics, and the accuracy of their critiques.

4.2.1 Critique Data Amount. Participants provided a total of 544 (mean = 17, SD = 5.82) comments as part of their critiques. These comments included feedback on the system, the data used in the explanation, and the presentation of the explanation. To gain a sense if participants' critique amount would change between explanations, we compared the number of comments they have made for each explanation they interacted with during the study.

Figure 2 (right) provides the average number of comments according to *Information Depth*. We saw a main effect of *Information Depth* on the number of comments provided ($F_{1,30} = 14.951$, $p < .001$), where participants provided more comments in the *detailed* explanation (mean = 9.84, SD = 4.31) in comparison to the *summary* explanation (mean = 7.16, SD = 2.46). This suggests that participants provided more thorough feedback on the system when presented with *detailed* explanations. The effect of *Progressive Disclosure* on the number of comments was not significant ($F_{1,30} = 0.943$, $p = .339$)

nor was the interaction effect of *Information Depth* × *Progressive Disclosure* ($F_{1,30} = 0.978$, $p = .33$)[1]. We further analyzed participants' average critique length (i.e., the average number of words in a comment) but we did not see any statistically significant effects ($p \geq .401$).

We also analyzed the distribution of participants' critiques regarding whether they primarily discussed strengths, weaknesses, or general critiques of the systems. With the *summary* explanation, participants identified 78 strengths,124 weaknesses, and 27 general comments, compared to 125 strengths, 151 weaknesses, and 39 general comments with the *detailed* explanation. Regarding *Progressive Disclosure*, participants identified 99 strengths, 123 weaknesses, 34 general comments when it was *present*, versus 104 strengths, 152 weaknesses, and 32 general comments when it was *absent*. The distribution of the comments did not differ significantly across *Information Depth* ($\chi^2(2, N = 544) = 2.17$, $p = .337$) or *Progressive Disclosure* ($\chi^2(2, N = 544) = 1.36$, $p = .506$).

4.2.2 Critique Data Coverage. We analyzed participants' critiques in terms of the coverage of topics presented in the explanation. For topic coverage, we label each comment in terms of the high-level categories presented in the explanation (Section 3.3). These categories include *Data Collection, Demographics, Usage,* and *General Information*. A few comments from participants focused on the overall dataset and did not fit under these categories. We labeled such comments under a "overall" category which resulted in five comment topics (*Data Collection, Demographics, Usage, General Information, Overall*).

At a group level, all the topics were covered in participants' critiques for both levels of *Information Depth*, supporting previous findings that participants collectively find value in all the presented information [6]. On an individual level, we saw a trend, concerning *Information Depth* ($F_{1,30} = 3.817$, $p = .06$) where participants covered a slightly higher number of topics with the *summary explanation* (mean = 3.78, SD = 1.18) than with the *detailed* explanation (mean = 3.28, SD = 0.99). However, as this effect only approaches significance ($p = .06$), we interpret this result as suggestive rather than definitive. There was no significant main effect of *Progressive Disclosure* ($F_{1,30} = 0.18$, $p = .674$) on the number of topics covered nor a significant *Information Depth* × *Progressive Disclosure* interaction ($F_{1,30} = 0.537$, $p = .469$).

We saw a statistically significant difference in the distribution of topics covered in participants' comments according to *Information Depth* ($\chi^2(4, N = 544) = 41.41$, $p < .00001$). Figure 3 (left) provides an overview of the distribution of topics and Table 3 provides some illustrative examples. In the *detailed* explanation, more than half of the comments focused on *Data Collection* whereas with *the summary* explanation, topic coverage was more balanced. This indicates that participants not only utilized the additional depth on the data collection process (e.g., specific details on collection tools and data pre-processing) in the *detailed* explanation, they also prioritized this information over other topics. Although additional details were available across multiple topics, participants did not use this information as much. This could either be because they found the *Data Collection* aspect to be the most critical or because the *Data Collection* information was presented as the first category. In contrast, with the *summary* explanations, participants spread their critiques more evenly across multiple topics. *Progressive Disclosure,* on the other hand, did not significantly impact the distribution of topics covered ($\chi^2(4, N = 544) = 5.35$, $p = .253$).

4.2.3 Critique Data Accuracy. We also coded the accuracy of the critique data to investigate if participants seemed to understand the information they were using as part of their critiques. To guide our coding, we followed prior work where participants' critique comments were coded in a 3-point scale (accurate, somewhat accurate, inaccurate) [6]. While we could label majority of the comments using this scale, some of the comments contained a question or stated uncertainty about the dataset. We labeled such comments as "uncertain".

Figure 3 (right) depicts a distribution of the critique accuracy across *Information Depth.* We found a statistically significant difference in the distribution of the critique accuracy across *Information Depth* ($\chi^2(3, N = 544) = 22.715$, $p < .0001$). Looking at the differences, we noticed that the participants provided slightly more accurate and somewhat accurate comments with the *detailed* explanation (accurate: 241/315, somewhat accurate: 58/315) than with the *summary* explanation (accurate: 153/229, somewhat accurate: 36/229). We further observed that participants showed more uncertainty in the *summary* explanation (34/229) than in the *detailed* explanation (15/315). Finally, we found only one comment in the *detailed* explanation to be completely inaccurate in comparison to the six comments in the *summary* explanation. The fact that only 7 out of 544 comments were rated as to be completely inaccurate suggests that both the explanations were at least moderately comprehensible for participants. This further supports earlier findings that participants can productively use training dataset explanation to generate reasonably accurate critique comments [6]. Here we demonstrated that participants can also achieve reasonably accurate critiques even with a summarized version. Table 4 provides some illustrative examples of critique comments in terms of their accuracy.

4.3 Impact of Information Depth and Progressive Disclosure on Comprehension Questionnaire

To gain an additional sense of participants' understanding of the explanation content, we analyzed responses to the two multiple choice questions and the three open-ended questions from the post-scenario comprehension questionnaire. We used a three-point scale to grade the open-ended questions (1 for correct answer, 0.5 for partially correct answer, 0 for incorrect answer). The average scores for both the *detailed* explanation (mean = 4.27, SD = 0.61) and *summary* explanation (mean = 4.06, SD = 0.79) were high, with no significant main effects of *Information Depth* ($F_{1,30} = 1.664$, $p = .207$) or *Progressive Disclosure* ($F_{1,30} = 1.484$, $p = .233$), and no *Information Depth* × *Progressive Disclosure* interaction effect ($F_{1,30} = 0.798$, $p = .37$). These questions were designed to be high-level to ensure they were answerable with both explanations, which likely contributed to the lack of significant effects. The questions and sample graded answers can be found in the auxiliary materials.

[1]For the cases where we used repeated measure ANOVA, we report the interaction effects as the test allows for this reporting.

Table 3: Illustrative Examples Of Participants' Critique Across Information Depth.

	Summary	Detailed
Collection	This is a large dataset which is helpful in developing accurate models. (P25)	Combination of manual and automated processes for data collection indicates a balanced approach that leverages both human expertise and technological efficiency in data collection. (P24)
Demographics	This transparency from demographics information enables the system to monitor and potentially mitigate biases in hiring decisions. (P23)	This does not reflect the general population, so you are training your system on non-representative data. This could lead to bias in its learning. (P32)
Usage	Clear guidance for usage ensures appropriate and ethical utilization, and helps prevent harm. (P24)	Good explanation on what this should be used for. (P13)
General Information	It is up to date and contains resumes received between 2015 and 2022. (P18)	No individual consent raises ethical issues and invasion of privacy. (P16)
Overall	Good summary to quickly explain but could use a bit more information regarding the AI. (P21)	Overall, the dataset, while might have its flaws, is great in assessing qualifications, experience and skills. (P6)

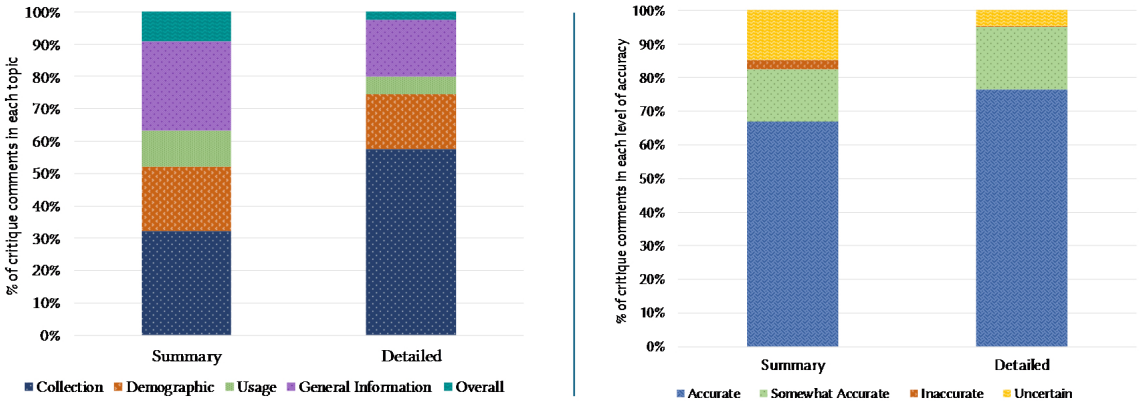


Figure 3: Left - Distributions of topics in participants' critique across Information Depth; Right - Distribution of the accuracy of participants' critique comments across Information Depth.

4.4 Interview Findings

We additionally examined our interview data to explore if they support or contradict our results from participants' critiques and questionnaire responses. Below we present our key findings from the interviews.

4.4.1 Unequivocal Preferences for Detailed Explanations. Our interview data suggests a clear preference for the *detailed* explanations, which supports our results from the questionnaire responses (Section 4.1) and participants' critique data (Section 4.2). Participants consistently emphasized the importance of *detailed* explanations to understand the systems' functionalities and implications.

> [I prefer the detailed] one because it gave me more information to make an informed decision if I'm going to purchase that system. But then the [summary] one, I would say, it was really vague. I didn't understand [it] fully, so I think I just felt more confident in the [detailed] one. (P5)

Most found the level of detail provided in the *detailed* explanation to be crucial for forming informed opinions about the AI systems and commented on how the lack of detail in the *summary* explanations led to frustration and uncertainty.

> I really felt like with the [summary] one, I didn't have enough context or enough information to really know how the data [was] being used or where it is coming from. So, the lack of information impacted my trust in the system. The [detailed] one, I feel like there was a lot more background information. (P25)

We saw a statistically significant impact of *Information Depth* on participants' perceived cognitive load (Section 4.1) and some participants in the interview also acknowledged the potential for information overload with the *detailed* explanation. However, they still preferred having the additional information to make informed decisions.

> To be honest, I didn't really have trouble, analyzing the [summary] one. I got confused at some point, but [that was] because there was not enough information. Even though I understood the [detailed one] and

Table 4: Illustrative examples for the critique accuracy ratings with researchers' interpretation of the critique comment in italics.

	Summary	Detailed
Accurate	This seems like a generally large sample, N=11k, which is good. (P33). *Comment on large sample size being helpful for training.*	Using average values for missing attributes might not provide accurate representation for individuals. (P20) *Comment on the concerns of using average values for missing data.*
Somewhat Accurate	Restricting the dataset's usage to only one specific purpose may limit its potential applications and insights that could benefit other areas. (P24) *Comment is somewhat correct in that it seemed to ignore the risk of using the same data in improper contexts without the usage guidance.*	This information is unnecessary. Knowing one's ethnicity could potentially introduce bias (both positive and negative). (P8) *Comment is somewhat correct as it seemed to ignore the fact that having such demographic information in the dataset might be necessary to understand the representativeness of the dataset.*
Incorrect	How would the individuals know their data was used? This information does not serve any purpose. (P18) *Participant incorrectly assume that the information on previous use cases is for the individuals whose data are in the dataset and not for the potential consumer of the dataset.*	If the data from applicants below 10 years old were used to train decision making for adult applicants, that would be questionable. (P5) *Participant misunderstood a part of the explanation that mentioned that 9% of the applicants came from the age group of over 55 years as applicants being less than 10 years old.*
Uncertain	This lacks information. Which 3 genders? Which 5 categories? How does ethnicity and country affect the application? (P13)	How do we know that this tool is reliable and accurate? What if it messed up data during the extraction process? (P11)

> would prefer it, it was a bit mentally challenging to analyze [the detailed explanation]. (P14)

4.4.2 Progressive Disclosure of Information can be a Useful Addition to Explanations. Participants, regardless of the presence of *Progressive Disclosure*, appreciated the well-organized interfaces with the mix of text and infographics. The majority of the participants who interacted with the explanation with *Progressive Disclosure* further commented on the interactive nature of the explanations, citing that it allowed them to control what information they accessed. This might help explain why participants reported learning more with *Progressive Disclosure.*

> When it comes to this [explanation] where you have to interact with the information, and you have to purposely [click] to look for the information, it kind of gives you a feeling of being in control of whatever information I'm seeing. This made me more interested in learning more in terms of the processes and reading more in depth in all the categories. (P28)

Other participants, however, did not find any potential benefit of *Progressive Disclosure,* particularly with the *summary* explanation.

> I felt like clicking sometimes was unnecessary because so little information [was revealed] in one click. (P1)

This indicates that *Progressive Disclosure* might be valued only when explanations contain substantial information.

5 Discussion

Our study showed that *Information Depth* strongly shapes users' perceptions of training dataset explanations, their critiques of the AI system that provide the explanation, and their cognitive load. *Detailed* explanation significantly increased participants' perceived understanding, trust, and fairness judgments of the AI system, as well as their perceived learning, though they also required more cognitive effort. Participants' critiques also reflected a tradeoff: *detailed* explanations led to more accurate critiques but narrower topic coverage (i.e., with a heavy focus on data collection information), while *summary* explanations prompted shorter critiques with slightly more balanced coverage across the different topics present in the explanation. *Progressive Disclosure* did not reduce cognitive load, but it did influence perceptions of learning. Participants expressed a consistent preference for *detailed* explanations despite the added effort.

5.1 Implications for Design

Our findings reveal several important design considerations for training dataset explanations. Participants' clear preference for *detailed* explanations despite the additional effort suggested that completeness and clarity may matter more than brevity when users seek to understand how an AI system was trained. This implies that explanation design should not default to minimalism; rather designers should prioritize presenting rich information in digestible ways.

Our results provide a starting point for establishing the role of Progressive Disclosure in explanation design. The fact that *Progressive Disclosure* did not reduce cognitive load, but enhanced participants' perception of learning suggests that its strength might lie less in lowering cognitive burden and more in shaping the user

experience, for example by supporting reflection and learning. Progressive Disclosure may be particularly useful in onboarding workflows, where supporting exploration, learning, and reflection is more important than immediate efficiency. Designers should therefore view Progressive Disclosure not as a universal solution for cognitive load management, but as one component in a broader design strategy that balances learning and usability.

The observed preference for *detailed* explanations emerged in a context where participants were presented with a reflective, high-stakes critique task. In such settings, users may be more willing to tolerate higher cognitive effort in exchange for a sense of completeness and understanding. This suggests that detailed explanations may be most appropriate during onboarding, auditing, or evaluation phases, when users are motivated to build a foundational mental model of the system. In contrast, users might benefit from *summary* explanations in lower-stakes or time constrained contexts. Designers should therefore treat explanation depth as context dependent choices, informed by users' goals and needs rather than fixed defaults.

Our findings highlight a potential tension between supporting both accurate and comprehensive critiques. Participants provided more critique comments with greater accuracy with the *detailed* explanation, but tended to focus predominantly on data collection (the first category in the interface), whereas explanation topic coverage was relatively more balanced with the *summary* explanation. Therefore, designers may need to consider the goal of the explanation when deciding on how depth to provide. For example, if the goal is to encourage users to consider all aspects of a system, and improving subjective impressions is not important, *summary* explanations might be most effective. On the other hand, *detailed* explanation might be better suited for critique accuracy and enhancing subjective impressions. Future work should explore explanations that are not only *detailed* but also incorporate mechanisms that motivate users to engage with different types of information. For example, one approach could be to highlight key pieces of information to draw attention [28, 51, 130]. Another approach could be to prompt users [117] to consider all relevant information.

5.2 Methodological and Theoretical Reflections

Contextualizing our results requires consideration of how *Information Depth* and *Progressive Disclosure* were implemented in the study. While participants' predominant focus on data collection information in their critique with the *detailed* explanation could indicate that they viewed this information as most critical, it is also possible that its placement at the top of the explanation biased their attention. Future research could examine how different information orders impact what users prioritize within the explanations. In addition, the longer critiques with the *detailed* explanations might simply reflect that there was more material to comment on. Future work should systematically investigate whether increased critique length reflects deeper understanding and more critical evaluation.

Regarding *Progressive Disclosure*, while the lack of impact on trust, fairness perception, and perceived understanding can be attributed to the fact that the same information was ultimately available across all conditions, it was surprising to see no differences in cognitive load based on suggestions in prior work [115, 136]. One plausible explanation is that the task context imposed only low to moderate cognitive load. Even the detailed explanations yielded relatively modest load (e.g., mean value of 3.5 on a 7-point scale), which might have left limited room for Progressive Disclosure to provide measurable relief. Additionally, given the absence of critical time pressure, participants may have been able to accommodate additional information without experiencing overload [57]. Progressive Disclosure might be more impactful in contexts characterized by time pressure or higher task complexity. It is also possible that alternative ways of achieving *Progressive Disclosure* might be more impactful, such as starting with a brief *summary* explanation and allowing users to expand it into a *detailed* version. Future research could also investigate other approaches to mitigating cognitive load, including cognitive forcing functions (e.g., mechanisms like checklists to slow down how people process information) [22, 37, 49], and alternative explanation formats (e.g., visuals, video, and hybrid approaches) [122].

Our study also raises questions about the role of prior knowledge of AI in shaping users' interactions with explanations. Although, we balanced participants' AI literacy [80, 126] and backgrounds across study conditions to minimize potential confounds associated with differing levels of AI knowledge, our participant pool demonstrated a relatively high mean AI literacy score (mean = 5.6, SD = 0.73), suggesting that many were generally knowledgeable about AI. Given recent findings that AI knowledge can shape user perceptions of AI explanations [43, 110, 122], future work should investigate whether AI background or literacy impacts how users leverage training dataset explanations. For example, it is possible that participants with lower AI literacy feel burdened by *detailed* explanations and prefer *summary* explanations.

Evaluating how well users comprehend explanations remains a methodological challenge. Given the limitation of depending solely on subjective measures [21, 54, 70, 101], we used critique-based measures and comprehension questionnaire alongside subjective ratings to better capture participants' understanding of the explanation and the system. Critiques revealed how participants acted on the information, while comprehension questionnaire provided an objective lens. However, assessing the accuracy of the critiques was challenging due to the variations in participants' interpretation of the same information and a lack of established ground truth (i.e., an objective standard used to evaluate the correctness of participants' critiques) [132]. Similarly, our comprehension questions were at a surface level to ensure that they were answerable with both explanations, limiting their ability to measure subtle differences in comprehension. Future work should explore how to design more nuanced comprehension questions that better differentiate between varying levels of understanding. Future research could also explore additional comprehension data collection techniques, such as think-aloud protocols [113] or eye-tracking [12, 87].

5.3 Study Limitations

Alongside the methodological and theoretical reflections noted in Section 5.2, we recognize the limitations of this study. First, our sample of 32 participants, a deliberate choice made to support manageable coding and in-depth qualitative analysis of critique data (Section 3.1), limited our statistical power. For Information Depth (a

within-subjects factor), the study was sufficiently powered to detect medium-to-large effects. In contrast, for Progressive Disclosure (a between-subjects factor), it was powered to detect only large effects. As such, the lack of significant findings for Progressive Disclosure should not be interpreted as evidence of no effect, but rather as an indication that smaller effects might not have been detectable with the current design. Future work should involve larger samples (e.g., around 300 participants), potentially using crowdsourcing, to better detect the smaller effects. Future work could also explore alternative study designs to increase the statistical power to detect differences across levels of *Progressive Disclosure* and to elicit contrastive comments on this factor.

Second, while our participants were generally motivated to engage with the explanations (a challenge identified in prior work [22, 52, 125]), the task of critiquing likely acted as a cognitive forcing function [72], encouraging deeper engagement with the explanations than what one might see in a more ecologically valid setting. Further, by explicitly framing participants as evaluators, the study may have introduced anchoring effects [123], leading them to calibrate their critiques around the richness of the information provided. As a result, this evaluative context may overestimate how much users would attend to detailed explanations in everyday use [137]. Future research should examine the generalizability of our findings to a broader variety of tasks, including those that do not inherently prompt critical thinking, such as simple information retrieval [83] or decision-support tasks [96].

Finally, although our scenario-based approach is consistent with prior work on AI explainability [5, 18, 41–43, 73, 118], it might not fully capture the complexities and nuances of real-world applications. Future studies should validate our findings in more real-word settings, particularly with domain experts [11, 55, 94, 122].

5.4 Broader Challenges for Effective Explanation Design

The increased trust and judgment of system fairness for *detailed* explanations over *summary* versions raises important considerations regarding the potential for overreliance [9, 99, 101, 128, 133]. The higher perceived trust and judgment of fairness suggests that users felt more confident in the system, however, this confidence is only beneficial for Human-AI collaboration if participants have engaged in critical thinking [62]. Using critiquing as a task in our study guarded against automatic thinking [62] and we observed positive impacts on critique accuracy with the *detailed* explanations. However, critique performance is still an indirect proxy for practical use, and our findings do not speak to whether participants could leverage the explanation content effectively in decision-making or appropriately calibrate their reliance on the AI system. This gap is especially important given prior work on the illusion of explanatory depth [34], where people overestimate their operational understanding [91]. Future work should therefore incorporate decision tasks with known ground truth and objective, performance-based measures such as decision accuracy [101], error detection [9] to provide a more complete account of the practical impact of training dataset explanations. An encouraging prior finding is that participants showed lower trust with detailed training dataset explanations with "red flags" compared to those that described fewer potential causes for concern [5], suggesting that depth can sometimes support calibration rather than inflate trust.

Despite their longer format than many AI explanations, the training dataset explanations as designed do not appear to suffer from the pitfall of high cognitive load. Training dataset explanations might still be susceptible to other explainability pitfalls outlined in Section 2.2, such as lack of actionability [45, 76] and misinterpretation [6]. Future research should explore how different explanation characteristics influence the number of actionable insights user derive from the explanation. Further studies should also explore how different presentation formats (e.g., textual, visual, or hybrid) [6, 122] influence the risk of misinterpretation or shape the kinds of insights users take away. Unless such pitfalls are addressed, transparency risks generating confusion rather than accountability.

6 Conclusion

In this paper, we examined how the depth of information in training dataset explanations and the use of Progressive Disclosure influence users' perceptions and understanding of an AI system and their cognitive load. In comparison to summary versions, detailed explanations improved users' understanding and subjective impressions of the system even though they led to higher cognitive load. Information depth further impacted the balance of topics that participants covered in their critiques of the system. While the use of Progressive Disclosure did not reduce cognitive load, it enhanced perceived learning. These findings highlight important design tradeoffs in training dataset explanations by emphasizing the need to balance completeness, clarity, and effort. Future research should further explore these tradeoffs, particularly in terms of how users' AI knowledge influences these results, or how the results might generalize to field studies. Additional work should also explore other design factors for training dataset explanations to improve user comprehension and satisfaction, thereby advancing the goal of achieving fully transparent and responsible human-centered AI systems.

Acknowledgments

We appreciate the valuable feedback from the anonymous reviewers, which helped improve the quality of this work. We also extend our gratitude to all study participants for their valuable time and contributions. We thank the Natural Sciences and Engineering Research Council of Canada (NSERC) and the University of Manitoba for their funding support.

GenAI Usage Disclosure

We used OpenAI's ChatGPT for grammar checking and text refinement only. All study design, prototypes, data analysis, interpretation of results were conducted solely by the authors. No GenAI tools were used to generate content, results, or references.

References

[1] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y. Lim, and Mohan Kankanhalli. 2018. Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (CHI '18), 1–18. https://doi.org/10.1145/3173574.3174156

[2] Mark S. Ackerman. 2000. The intellectual challenge of CSCW: the gap between social requirements and technical feasibility. *Hum.-Comput. Interact.* 15, 2: 179–203. https://doi.org/10.1207/S15327051HCI1523_5

[3] Amina Adadi and Mohammed Berrada. 2018. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access* 6: 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

[4] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2022. Machine Bias *. In *Ethics of Data and Analytics*. Auerbach Publications.

[5] Ariful Islam Anik and Andrea Bunt. 2021. Data-Centric Explanations: Explaining Training Data of Machine Learning Systems to Promote Transparency. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (CHI '21). https://doi.org/10.1145/3411764.3445736

[6] Ariful Islam Anik and Andrea Bunt. 2024. Supporting User Critiques of AI Systems via Training Dataset Explanations: Investigating Critique Properties and the Impact of Presentation Style. In *2024 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*. https://doi.org/10.1109/VL/HCC60511.2024.00024

[7] Vijay Arya, Rachel KE Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C Hoffman, Stephanie Houde, Q Vera Liao, Ronny Luss, Aleksandra Mojsilović, and others. 2019. One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques. *arXiv preprint arXiv:1909.03012*.

[8] Maria Avgerinou and John Ericson. 1997. A Review of the Concept of Visual Literacy. *British Journal of Educational Technology* 28, 4: 280–291. https://doi.org/10.1111/1467-8535.00035

[9] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (CHI '21), 1–16. https://doi.org/10.1145/3411764.3445717

[10] Gustavo E. A. P. A. Batista, Ronaldo C. Prati, and Maria Carolina Monard. 2004. A study of the behavior of several methods for balancing machine learning training data. *SIGKDD Explor. Newsl.* 6, 1: 20–29. https://doi.org/10.1145/1007730.1007735

[11] Sarah Bayer, Henner Gimpel, and Moritz Markgraf. 2022. The role of domain expertise in trusting and following explainable AI decision support systems. *Journal of Decision Systems* 32, 1: 110–138. https://doi.org/10.1080/12460125.2021.1958505

[12] Roman Bednarik and Markku Tukiainen. 2006. An eye-tracking methodology for characterizing program comprehension processes. In *Proceedings of the 2006 symposium on Eye tracking research & applications* (ETRA '06), 125–132. https://doi.org/10.1145/1117309.1117356

[13] Emily M. Bender and Batya Friedman. 2018. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics* 6: 587–604. https://doi.org/10.1162/tacl_a_00041

[14] Mostafa Ali Benzaghta, Abdulaziz Elwalda, Mousa Mousa, Ismail Erkan, and Mushfiqur Rahman. 2021. SWOT analysis applications: An integrative literature review. *Journal of Global Business Insights* 6, 1: 55–73. https://doi.org/10.5038/2640-6489.6.1.1148

[15] Niels van Berkel, Jorge Goncalves, Daniel Russo, Simo Hosio, and Mikael B. Skov. 2021. Effect of Information Presentation on Fairness Perceptions of Machine Learning Predictors. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (CHI '21), 1–13. https://doi.org/10.1145/3411764.3445365

[16] Aditya Bhattacharya, Jeroen Ooge, Gregor Stiglic, and Katrien Verbert. 2023. Directive Explanations for Monitoring the Risk of Diabetes Onset: Introducing Directive Data-Centric Explanations and Combinations to Support What-If Explorations. In *Proceedings of the 28th International Conference on Intelligent User Interfaces* (IUI '23), 204–219. https://doi.org/10.1145/3581641.3584075

[17] Aditya Bhattacharya, Simone Stumpf, Lucija Gosak, Gregor Stiglic, and Katrien Verbert. 2024. EXMOS: Explanatory Model Steering through Multifaceted Explanations and Data Configurations. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (CHI '24), 1–27. https://doi.org/10.1145/3613904.3642106

[18] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. "It's Reducing a Human Being to a Percentage": Perceptions of Justice in Algorithmic Decisions. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3173574.3173951

[19] Sarah Brayne. 2020. *Predict and Surveil: Data, Discretion, and the Future of Policing*. Oxford University Press.

[20] Sarah Brayne and Angèle Christin. 2021. Technologies of Crime Prediction: The Reception of Algorithms in Policing and Criminal Courts. *Social Problems* 68, 3: 608–624. https://doi.org/10.1093/socpro/spaa004

[21] Zana Buçinca, Phoebe Lin, Krzysztof Z. Gajos, and Elena L. Glassman. 2020. Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces* (IUI '20), 454–464. https://doi.org/10.1145/3377325.3377498

[22] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. 2021. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1: 188:1-188:21. https://doi.org/10.1145/3449287

[23] Adrian Bussone, Simone Stumpf, and Dympna O'Sullivan. 2015. The Role of Explanations on Trust and Reliance in Clinical Decision Support Systems. In *2015 International Conference on Healthcare Informatics*, 160–169. https://doi.org/10.1109/ICHI.2015.26

[24] Carrie J. Cai, Jonas Jongejan, and Jess Holbrook. 2019. The effects of example-based explanations in a machine learning interface. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, 258–262. https://doi.org/10.1145/3301275.3302289

[25] Carrie J. Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2019. "Hello AI": Uncovering the Onboarding Needs of Medical Practitioners for Human-AI Collaborative Decision-Making. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW: 104:1-104:24. https://doi.org/10.1145/3359206

[26] Kelly Caine. 2016. Local Standards for Sample Size at CHI. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 981–992. https://doi.org/10.1145/2858036.2858498

[27] Paul Cairns. 2016. 34. Experimental Methods in Human-Computer Interaction.

[28] Giuseppe Carenini, Cristina Conati, Enamul Hoque, Ben Steichen, Dereck Toker, and James Enns. 2014. Highlighting interventions and user differences: informing adaptive information visualization support. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1835–1844. https://doi.org/10.1145/2556288.2557141

[29] John M. Carroll and Caroline Carrithers. 1984. Training wheels in a user interface. *Communications of the ACM* 27, 8: 800–806. https://doi.org/10.1145/358198.358218

[30] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. 2015. Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1721–1730. https://doi.org/10.1145/2783258.2788613

[31] Mohamed Amine Chatti, Mouadh Guesmi, Laura Vorgerd, Thao Ngo, Shoeb Joarder, Qurat Ul Ain, and Arham Muslim. 2022. Is More Always Better? The Effects of Personal Characteristics and Level of Detail on the Perception of Explanations in a Recommender System. In *Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization* (UMAP '22), 254–264. https://doi.org/10.1145/3503252.3531304

[32] Hao-Fei Cheng, Ruotong Wang, Zheng Zhang, Fiona O'Connell, Terrance Gray, F. Maxwell Harper, and Haiyi Zhu. 2019. Explaining Decision-Making Algorithms through UI: Strategies to Help Non-Expert Stakeholders. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (CHI '19), 1–12. https://doi.org/10.1145/3290605.3300789

[33] Michael Chromik and Andreas Butz. 2021. Human-XAI Interaction: A Review and Design Principles for Explanation User Interfaces. In *Human-Computer Interaction – INTERACT 2021*, Carmelo Ardito, Rosa Lanzilotti, Alessio Malizia, Helen Petrie, Antonio Piccinno, Giuseppe Desolda and Kori Inkpen (eds.). Springer International Publishing, Cham, 619–640. https://doi.org/10.1007/978-3-030-85616-8_36

[34] Michael Chromik, Malin Eiband, Felicitas Buchner, Adrian Krüger, and Andreas Butz. 2021. I Think I Get Your Point, AI! The Illusion of Explanatory Depth in Explainable AI. In *Proceedings of the 26th International Conference on Intelligent User Interfaces* (IUI '21), 307–317. https://doi.org/10.1145/3397481.3450644

[35] Henriette Cramer, Vanessa Evers, Satyan Ramlal, Maarten Van Someren, Lloyd Rutledge, Natalia Stash, Lora Aroyo, and Bob Wielinga. 2008. The effects of transparency on trust in and acceptance of a content-based art recommender. *User Modeling and User-Adapted Interaction* 18, 5: 455–496. https://doi.org/10.1007/s11257-008-9051-3

[36] Anamaria Crisan, Margaret Drouhard, Jesse Vig, and Nazneen Rajani. 2022. Interactive Model Cards: A Human-Centered Approach to Model Documentation. In *2022 ACM Conference on Fairness Accountability and Transparency*, 427–439. https://doi.org/10.1145/3531146.3533108

[37] Pat Croskerry, Geeta Singhal, and Sílvia Mamede. 2013. Cognitive debiasing 2: impediments to and strategies for change. *BMJ quality & safety* 22 Suppl 2, Suppl 2: ii65–ii72. https://doi.org/10.1136/bmjqs-2012-001713

[38] Anupam Datta, Shayak Sen, and Yair Zick. 2016. Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems. In *2016 IEEE Symposium on Security and Privacy (SP)*, 598–617. https://doi.org/10.1109/SP.2016.42

[39] Jonathan Dodge, Q. Vera Liao, Yunfeng Zhang, Rachel K. E. Bellamy, and Casey Dugan. 2019. Explaining models: an empirical study of how explanations impact fairness judgment. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (IUI '19), 275–285. https://doi.org/10.1145/3301275.3302310

[40] Finale Doshi-Velez and Been Kim. 2017. Towards A Rigorous Science of Interpretable Machine Learning. https://doi.org/10.48550/arXiv.1702.08608

[41] Upol Ehsan, Q. Vera Liao, Michael Muller, Mark O. Riedl, and Justin D. Weisz. 2021. Expanding Explainability: Towards Social Transparency in AI systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (CHI '21), 1–19. https://doi.org/10.1145/3411764.3445188

[42] Upol Ehsan, Q. Vera Liao, Samir Passi, Mark O. Riedl, and Hal Daumé. 2024. Seamful XAI: Operationalizing Seamful Design in Explainable AI. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW1: 1–29. https://doi.org/10.

1145/3637396
[43] Upol Ehsan, Samir Passi, Q. Vera Liao, Larry Chan, I-Hsiang Lee, Michael Muller, and Mark O Riedl. 2024. The Who in XAI: How AI Background Shapes Perceptions of AI Explanations. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (CHI '24), 1–32. https://doi.org/10.1145/3613904.3642474
[44] Upol Ehsan and Mark O. Riedl. 2020. Human-Centered Explainable AI: Towards a Reflective Sociotechnical Approach. In *HCI International 2020 - Late Breaking Papers: Multimodality and Intelligence*, 449–466. https://doi.org/10.1007/978-3-030-60117-1_33
[45] Upol Ehsan and Mark O. Riedl. 2024. Explainability pitfalls: Beyond dark patterns in explainable AI. *Patterns* 5, 6. https://doi.org/10.1016/j.patter.2024.100971
[46] Upol Ehsan, Koustuv Saha, Munmun De Choudhury, and Mark O. Riedl. 2023. Charting the Sociotechnical Gap in Explainable AI: A Framework to Address the Gap in XAI. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1: 1–32. https://doi.org/10.1145/3579467
[47] Upol Ehsan, Philipp Wintersberger, Q. Vera Liao, Elizabeth Anne Watkins, Carina Manger, Hal Daumé III, Andreas Riener, and Mark O Riedl. 2022. Human-Centered Explainable AI (HCXAI): Beyond Opening the Black-Box of AI. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems* (CHI EA '22), 1–7. https://doi.org/10.1145/3491101.3503727
[48] Malin Eiband, Hanna Schneider, Mark Bilandzic, Julian Fazekas-Con, Mareike Haug, and Heinrich Hussmann. 2018. Bringing Transparency Design into Practice. In *23rd International Conference on Intelligent User Interfaces*, 211–223. https://doi.org/10.1145/3172944.3172961
[49] John W. Ely, Mark L. Graber, and Pat Croskerry. 2011. Checklists to Reduce Diagnostic Errors. *Academic Medicine* 86, 3: 307. https://doi.org/10.1097/ACM.0b013e31820824cd
[50] Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun. 2017. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 542, 7639: 115–118. https://doi.org/10.1038/nature21056
[51] Robert L. Fowler and Anne S. Barker. 1974. Effectiveness of highlighting for retention of text material. *Journal of Applied Psychology* 59, 3: 358–364. https://doi.org/10.1037/h0036750
[52] Krzysztof Z. Gajos and Lena Mamykina. 2022. Do People Engage Cognitively with AI? Impact of AI Assistance on Incidental Learning. In *27th International Conference on Intelligent User Interfaces*, 794–806. https://doi.org/10.1145/3490099.3511138
[53] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM* 64, 12: 86–92. https://doi.org/10.1145/3458723
[54] Ben Green and Yiling Chen. 2019. Disparate Interactions: An Algorithm-in-the-Loop Analysis of Fairness in Risk Assessments. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 90–99. https://doi.org/10.1145/3287560.3287563
[55] Shirley Gregor and Izak Benbasat. 1999. Explanations from Intelligent Systems: Theoretical Foundations and Implications for Practice. *MIS Quarterly* 23, 4: 497–530. https://doi.org/10.2307/249487
[56] David Gunning and David Aha. 2019. DARPA's explainable artificial intelligence (XAI) program. *AI magazine* 40, 2: 44–58.
[57] Minhi Hahn, Robert Lawson, and Young Gyu Lee. 1992. The effects of time pressure and information load on decision quality. *Psychology & Marketing* 9, 5: 365–378. https://doi.org/10.1002/mar.4220090503
[58] Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Advances in Psychology*. Elsevier, 139–183. https://doi.org/10.1016/S0166-4115(08)62386-9
[59] Marilyn M. Helms and Judy Nixon. 2010. Exploring SWOT analysis – where are we now?: A review of academic research from the last decade. *Journal of Strategy and Management* 3, 3: 215–251. https://doi.org/10.1108/17554251011064837
[60] Fred Hohman, Andrew Head, Rich Caruana, Robert DeLine, and Steven M. Drucker. 2019. Gamut: A Design Probe to Understand How Data Scientists Understand Machine Learning Models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (CHI '19), 1–13. https://doi.org/10.1145/3290605.3300809
[61] Hsiu-Fang Hsieh and Sarah E. Shannon. 2005. Three Approaches to Qualitative Content Analysis. *Qualitative Health Research* 15, 9: 1277–1288. https://doi.org/10.1177/1049732305276687
[62] Daniel Kahneman. 2011. *Thinking, fast and slow*. Farrar, Straus and Giroux, New York, NY, US.
[63] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. 2020. Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (CHI '20), 1–14. https://doi.org/10.1145/3313831.3376219
[64] René F. Kizilcec. 2016. How Much Information? Effects of Transparency on Trust in an Algorithmic Interface. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (CHI '16), 2390–2395. https://doi.org/10.1145/2858036.2858402
[65] Joshua Klayman. 1995. Varieties of Confirmation Bias. In *Psychology of Learning and Motivation*, Jerome Busemeyer, Reid Hastie and Douglas L. Medin (eds.). Academic Press, 385–418. https://doi.org/10.1016/S0079-7421(08)60315-1
[66] Rafal Kocielnik, Saleema Amershi, and Paul N. Bennett. 2019. Will You Accept an Imperfect AI? Exploring Designs for Adjusting End-user Expectations of AI Systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (CHI '19), 1–14. https://doi.org/10.1145/3290605.3300641
[67] Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. 2015. Principles of Explanatory Debugging to Personalize Interactive Machine Learning. In *Proceedings of the 20th International Conference on Intelligent User Interfaces*, 126–137. https://doi.org/10.1145/2678025.2701399
[68] Todd Kulesza, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan, and Weng-Keen Wong. 2013. Too much, too little, or just right? Ways explanations impact end users' mental models. In *2013 IEEE Symposium on Visual Languages and Human Centric Computing*, 3–10. https://doi.org/10.1109/VLHCC.2013.6645235
[69] Nathan R. Kuncel, Deniz S. Ones, and David M. Klieger. 2014. In Hiring, Algorithms Beat Instinct. *Harvard Business Review*. Retrieved August 26, 2024 from https://hbr.org/2014/05/in-hiring-algorithms-beat-instinct
[70] Vivian Lai and Chenhao Tan. 2019. On Human Predictions with Explanations and Predictions of Machine Learning Models: A Case Study on Deception Detection. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (FAT* '19), 29–38. https://doi.org/10.1145/3287560.3287590
[71] Vivian Lai, Yiming Zhang, Chacha Chen, Q. Vera Liao, and Chenhao Tan. 2023. Selective Explanations: Leveraging Human Input to Align Explainable AI. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW2: 357:1-357:35. https://doi.org/10.1145/3610206
[72] Kathryn Ann Lambe, Gary O'Reilly, Brendan D. Kelly, and Sarah Curristan. 2016. Dual-process cognitive interventions to enhance diagnostic reasoning: a systematic review. *BMJ Quality & Safety* 25, 10: 808–820. https://doi.org/10.1136/bmjqs-2015-004417
[73] Min Kyung Lee. 2018. Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society* 5, 1: 2053951718756684. https://doi.org/10.1177/2053951718756684
[74] Q. Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (CHI '20), 1–15. https://doi.org/10.1145/3313831.3376590
[75] Q. Vera Liao and Kush R. Varshney. 2022. Human-Centered Explainable AI (XAI): From Algorithms to User Experiences. Retrieved August 21, 2024 from http://arxiv.org/abs/2110.10790
[76] Q. Vera Liao, Yunfeng Zhang, Ronny Luss, Finale Doshi-Velez, and Amit Dhurandhar. 2022. Connecting Algorithmic Research and Usage Contexts: A Perspective of Contextualized Evaluation for Explainable AI. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 10: 147–159. https://doi.org/10.1609/hcomp.v10i1.21995
[77] Brian Y. Lim, Anind K. Dey, and Daniel Avrahami. 2009. Why and why not explanations improve the intelligibility of context-aware intelligent systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '09), 2119–2128. https://doi.org/10.1145/1518701.1519023
[78] Tania Lombrozo. 2007. Simplicity and probability in causal explanation. *Cognitive Psychology* 55, 3: 232–257. https://doi.org/10.1016/j.cogpsych.2006.09.006
[79] Tania Lombrozo. 2016. Explanatory Preferences Shape Learning and Inference. *Trends in Cognitive Sciences* 20, 10: 748–759. https://doi.org/10.1016/j.tics.2016.08.001
[80] Duri Long and Brian Magerko. 2020. What is AI Literacy? Competencies and Design Considerations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (CHI '20), 1–16. https://doi.org/10.1145/3313831.3376727
[81] Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems*. Retrieved from https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf
[82] Gideon Mann and Cathy O'Neil. 2016. Hiring algorithms are not neutral. *Harvard Business Review* 9: 2016.
[83] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to information retrieval*. Cambridge university press, Cambridge.
[84] Mark Mazumder, Colby Banbury, Xiaozhe Yao, Bojan Karlaš, William Gaviria Rojas, Sudnya Diamos, Greg Diamos, Lynn He, Alicia Parrish, Hannah Rose Kirk, Jessica Quaye, Charvi Rastogi, Douwe Kiela, David Jurado, David Kanter, Rafael Mosquera, Juan Ciro, Lora Aroyo, Bilge Acun, Lingjiao Chen, Mehul Smriti Raje, Max Bartolo, Sabri Eyuboglu, Amirata Ghorbani, Emmett Goodman, Oana Inel, Tariq Kane, Christine R Kirkpatrick, Tzu-Sheng Kuo, Jonas Mueller, Tristan Thrush, Joaquin Vanschoren, Margaret Warren, Adina Williams, Serena Yeung, Newsha Ardalani, Praveen Paritosh, Ce Zhang, James Zou, Carole-Jean Wu, Cody Coleman, Andrew Ng, Peter Mattson, and Vijay Janapa Reddi. DataPerf: Benchmarks for Data-Centric AI Development.

[85] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. 2019. Reliability and Inter-rater Reliability in Qualitative Research: Norms and Guidelines for CSCW and HCI Practice. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW: 1–23. https://doi.org/10.1145/3359174

[86] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A Survey on Bias and Fairness in Machine Learning. *ACM Comput. Surv.* 54, 6: 115:1-115:35. https://doi.org/10.1145/3457607

[87] Diane C. Mézière, Lili Yu, Erik D. Reichle, Titus von der Malsburg, and Genevieve McArthur. 2023. Using Eye-Tracking Measures to Predict Reading Comprehension. *Reading Research Quarterly* 58, 3: 425–449. https://doi.org/10.1002/rrq.498

[88] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267: 1–38. https://doi.org/10.1016/j.artint.2018.07.007

[89] Tim Miller, Piers Howe, and Liz Sonenberg. 2017. Explainable AI: Beware of Inmates Running the Asylum Or: How I Learnt to Stop Worrying and Love the Social and Behavioural Sciences. https://doi.org/10.48550/arXiv.1712.00547

[90] 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. https://doi.org/10.1145/3287560.3287596

[91] Shane T Mueller, Robert R Hoffman, William Clancey, Abigail Emrey, and Gary Klein. 2019. Explanation in human-AI systems: A literature meta-review, synopsis of key ideas and publications, and bibliography for explainable AI. *arXiv preprint arXiv:1902.01876*.

[92] Lloyd H. Nakatani and John A. Rohrlich. 1983. Soft machines: A philosophy of user-computer interface design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '83), 19–23. https://doi.org/10.1145/800045.801573

[93] Donald A. Norman. 2002. *The Design of Everyday Things*. Basic Books, Inc., USA.

[94] Mahsan Nourani, Joanie King, and Eric Ragan. 2020. The Role of Domain Expertise in User Trust and the Impact of First Impressions with Intelligent Systems. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 8: 112–121. https://doi.org/10.1609/hcomp.v8i1.7469

[95] Conor Nugent and Pádraig Cunningham. 2005. A Case-Based Explanation System for Black-Box Systems. *Artificial Intelligence Review* 24, 2: 163–178. https://doi.org/10.1007/s10462-005-4609-5

[96] Ingrid Nunes and Dietmar Jannach. 2017. A systematic review and taxonomy of explanations in decision support and recommender systems. *User Modeling and User-Adapted Interaction* 27, 3–5: 393–444. https://doi.org/10.1007/s11257-017-9195-0

[97] Fred Paas, Alexander Renkl, and John Sweller. 2003. Cognitive Load Theory and Instructional Design: Recent Developments. *Educational Psychologist* 38, 1: 1–4. https://doi.org/10.1207/S15326985EP3801_1

[98] Frank Pasquale. 2015. *The Black Box Society: The Secret Algorithms That Control Money and Information*. Harvard University Press. https://doi.org/10.4159/harvard.9780674736061

[99] Samir Passi and Mihaela Vorvoreanu. Overreliance on AI Literature Review.

[100] John W Payne, James R Bettman, and Eric J Johnson. 1988. Adaptive strategy selection in decision making. *Journal of experimental psychology: Learning, Memory, and Cognition* 14, 3: 534.

[101] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and Measuring Model Interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (CHI '21), 1–52. https://doi.org/10.1145/3411764.3445315

[102] Mahima Pushkarna, Andrew Zaldivar, and Oddur Kjartansson. 2022. Data Cards: Purposeful and Transparent Dataset Documentation for Responsible AI. In *2022 ACM Conference on Fairness Accountability and Transparency*, 1776–1826. https://doi.org/10.1145/3531146.3533231

[103] Emilee Rader, Kelley Cotter, and Janghee Cho. 2018. Explanations as Mechanisms for Supporting Algorithmic Transparency. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3173574.3173677

[104] McKenzie Raub. Bots, Bias and Big Data: Artificial Intelligence, Algorithmic Bias and Disparate Impact Liability in Hiring Practices. *ARKANSAS LAW REVIEW* 71.

[105] Raymond Reiter. 1987. A theory of diagnosis from first principles. *Artificial Intelligence* 32, 1: 57–95. https://doi.org/10.1016/0004-3702(87)90062-2

[106] Marco Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, 97–101. https://doi.org/10.18653/v1/N16-3020

[107] Vincent Robbemond, Oana Inel, and Ujwal Gadiraju. 2022. Understanding the Role of Explanation Modality in AI-assisted Decision-making. In *Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization*, 223–233. https://doi.org/10.1145/3503252.3531311

[108] Mary Beth Rosson and John M. Carroll. 2012. Scenario-Based Design. In *Human Computer Interaction Handbook* (3rd ed.). CRC Press.

[109] Wojciech Samek, Alexander Binder, Gregoire Montavon, Sebastian Lapuschkin, and Klaus-Robert Muller. 2017. Evaluating the Visualization of What a Deep Neural Network Has Learned. *IEEE Transactions on Neural Networks and Learning Systems* 28, 11: 2660–2673. https://doi.org/10.1109/TNNLS.2016.2599820

[110] James Schaffer, John O'Donovan, James Michaelis, Adrienne Raglin, and Tobias Höllerer. 2019. I can do better than your AI: expertise and explanations. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (IUI '19), 240–251. https://doi.org/10.1145/3301275.3302308

[111] Jakob Schoeffer, Niklas Kuehl, and Yvette Machowski. 2022. "There Is Not Enough Information": On the Effects of Explanations on Perceptions of Informational Fairness and Trustworthiness in Automated Decision-Making. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (FAccT '22), 1616–1628. https://doi.org/10.1145/3531146.3533218

[112] Jane Secker and Emma Coonan (eds.). 2012. *Rethinking Information Literacy: A Practical Framework for Supporting Learning*. Facet. https://doi.org/10.29085/9781856049528

[113] Maarten W. van Someren and And Others. 1994. *The Think Aloud Method: A Practical Guide to Modelling Cognitive Processes*. Academic Press, Inc.

[114] Frank Spillers. 2004. What is Progressive Disclosure? - ED. *Experience Dynamics*. Retrieved August 21, 2024 from https://www.experiencedynamics.com/progressive-disclosure-the-best-interaction-design-technique/

[115] Aaron Springer and Steve Whittaker. 2020. Progressive Disclosure: When, Why, and How Do Users Want Algorithmic Transparency Information? *ACM Transactions on Interactive Intelligent Systems* 10, 4: 1–32. https://doi.org/10.1145/3374218

[116] Centre for Education Statistics and Evaluation. 2017. Cognitive load theory: Research that teachers really need to understand. *Sydney: Centre for Education Statistics and Evaluation*.

[117] Teresa M. Stephens. 2020. Cognitive Debiasing: *Nurse Leader* 18, 4: 344–351. https://doi.org/10.1016/j.mnl.2019.03.013

[118] Jiao Sun, Q. Vera Liao, Michael Muller, Mayank Agarwal, Stephanie Houde, Kartik Talamadupula, and Justin D. Weisz. 2022. Investigating Explainability of Generative AI for Code through Scenario-based Design. In *Proceedings of the 27th International Conference on Intelligent User Interfaces* (IUI '22), 212–228. https://doi.org/10.1145/3490099.3511119

[119] John Sweller. 2010. Element Interactivity and Intrinsic, Extraneous, and Germane Cognitive Load. *Educational Psychology Review* 22, 2: 123–138. https://doi.org/10.1007/s10648-010-9128-5

[120] John Sweller. 2011. Cognitive Load Theory. In *Psychology of Learning and Motivation*. Elsevier, 37–76. https://doi.org/10.1016/B978-0-12-387691-1.00002-8

[121] John Sweller, Jeroen J. G. Van Merrienboer, and Fred G. W. C. Paas. 1998. Cognitive Architecture and Instructional Design. *Educational Psychology Review* 10, 3: 251–296. https://doi.org/10.1023/A:1022193728205

[122] Maxwell Szymanski, Martijn Millecamp, and Katrien Verbert. 2021. Visual, textual or hybrid: the effect of user expertise on different explanations. In *Proceedings of the 26th International Conference on Intelligent User Interfaces* (IUI '21), 109–119. https://doi.org/10.1145/3397481.3450662

[123] Amos Tversky and Daniel Kahneman. 1974. Judgment under Uncertainty: Heuristics and Biases. *Science* 185, 4157: 1124–1131. https://doi.org/10.1126/science.185.4157.1124

[124] Jeroen J. G. Van Merriënboer and John Sweller. 2005. Cognitive Load Theory and Complex Learning: Recent Developments and Future Directions. *Educational Psychology Review* 17, 2: 147–177. https://doi.org/10.1007/s10648-005-3951-0

[125] Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael S. Bernstein, and Ranjay Krishna. 2023. Explanations Can Reduce Overreliance on AI Systems During Decision-Making. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1: 1–38. https://doi.org/10.1145/3579605

[126] Bingcheng Wang, Pei-Luen Patrick Rau, and Tianyi Yuan. 2023. Measuring user competence in using artificial intelligence: validity and reliability of artificial intelligence literacy scale. *Behaviour & Information Technology* 42, 9: 1324–1337. https://doi.org/10.1080/0144929X.2022.2072768

[127] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y. Lim. 2019. Designing Theory-Driven User-Centric Explainable AI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (CHI '19), 1–15. https://doi.org/10.1145/3290605.3300831

[128] Xinru Wang and Ming Yin. 2021. Are Explanations Helpful? A Comparative Study of the Effects of Explanations in AI-Assisted Decision-Making. In *26th International Conference on Intelligent User Interfaces*, 318–328. https://doi.org/10.1145/3397481.3450650

[129] Daniel S. Weld and Gagan Bansal. 2019. The challenge of crafting intelligible intelligence. *Commun. ACM* 62, 6: 70–79. https://doi.org/10.1145/3282486

[130] Jen-Her Wu and Yufei Yuan. 2003. Improving searching and reading performance: the effect of highlighting and text color coding. *Inf. Manage.* 40, 7: 617–637. https://doi.org/10.1016/S0378-7206(02)00091-5

[131] L. Richard Ye and Paul E. Johnson. 1995. The Impact of Explanation Facilities on User Acceptance of Expert Systems Advice. *MIS Quarterly* 19, 2: 157–172.

https://doi.org/10.2307/249686

[132] Hubert Dariusz Zając, Natalia Rozalia Avlona, Finn Kensing, Tariq Osman Andersen, and Irina Shklovski. 2023. Ground Truth Or Dare: Factors Affecting The Creation Of Medical Datasets For Training AI. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society* (AIES '23), 351–362. https://doi.org/10.1145/3600211.3604766

[133] Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 295–305. https://doi.org/10.1145/3351095.3372852

[134] 2024. PAIR-code/knowyourdata. Retrieved August 26, 2024 from https://github.com/PAIR-code/knowyourdata

[135] 2024. Userback: Your #1 User Feedback Software. Retrieved August 1, 2024 from https://userback.io/

[136] Progressive Disclosure. *Nielsen Norman Group*. Retrieved December 9, 2024 from https://www.nngroup.com/articles/progressive-disclosure/

[137] The Hawthorne Effect or Observer Bias in User Research. *Nielsen Norman Group*. Retrieved December 15, 2025 from https://www.nngroup.com/articles/hawthorne-effect-observer-bias-user-research/