From Curiosity to Caution: How Expertise Shapes the Use of Interpretable Machine Learning

SYEDA MASOOMA NAQVI, University of Minnesota, USA
GAYATHRI BALAJI, University of Minnesota, USA
ANGELIQUE LAJU, University of Minnesota, USA
HARMANPREET KAUR, University of Minnesota, USA

Interpretability tools are increasingly used to make ML models more transparent, yet their effectiveness has been limited in practice despite significant work on improving their design. While individual factors (e.g., mental models, cognitive biases) shape tool use, the research community has also highlighted the influence of socio-organizational factors (e.g., job roles, team dynamics, policies). We trace the impact of one such factor that operates at a hierarchical level: proficiency differential, i.e., the development of skill over time as people evolve from novices to experts. To investigate the value of this proficiency differential, we conducted contextual inquiries and semi-structured interviews with expert and novice data scientists (N=23). Our work contributes empirical evidence of how proficiency shapes interpretability use, with novices driven by curiosity and experts by efficiency and caution. We present a framework for understanding expert–novice differences, and identify design implications for interpretable ML that scaffold both expert-like reasoning and novice-like exploration.

CCS Concepts: • **Human-centered computing** → **Empirical studies in HCI**; • **Computing methodologies** → *Machine learning*.

Additional Key Words and Phrases: Interpretability, Machine Learning, Expertise, Practitioners

ACM Reference Format:
Syeda Masooma Naqvi, Gayathri Balaji, Angelique Laju, and Harmanpreet Kaur. 2026. From Curiosity to Caution: How Expertise Shapes the Use of Interpretable Machine Learning. In *The 2026 ACM Conference on Fairness, Accountability, and Transparency (FAccT '26), June 25–28, 2026, Montreal, QC, Canada.* ACM, New York, NY, USA, 29 pages. https://doi.org/10.1145/3805689.3812376

1 Introduction

The growing accessibility and wide range of applications of data science techniques have made them an integral part of many people's workflows. As we collect more and more data in our information ecosystem, data science is now routinely applied in diverse domains—from weather forecasting [163] and sports analytics [131] to social media content filtering [5], and more sensitive settings such as education [124], healthcare [82], and finance [111, 153]. Moreover, with the democratization and accessibility of data science techniques and tools, people with varying levels of expertise can effectively apply these methods. However, with these rising applications and access to data science, model transparency and accountability have become critical concerns. Many ML models, after being applied in practice, are discovered to be biased, imbalanced, missing relevant edge cases and guardrails, and riddled with issues that can cause harm (e.g., [7, 54, 78, 94, 128, 151, 155, 181, 188]).

Authors' Contact Information: Syeda Masooma Naqvi, naqvi042@umn.edu, University of Minnesota, Minneapolis, USA; Gayathri Balaji, balaj069@umn.edu, University of Minnesota, Minneapolis, USA; Angelique Laju, laju0002@umn.edu, University of Minnesota, Minneapolis, USA; Harmanpreet Kaur, harmank@umn.edu, University of Minnesota, Minneapolis, USA.

FAccT '26, Montreal, QC, Canada

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/10.1145/3805689.3812376

Interpretability and explainability tools[1] are proposed as a solution to support model transparency and accountability [60, 62, 126], but have had mixed success in practice. Prior work has studied user reliance on interpretability tools for tasks such as exploratory data science (e.g., [22, 79, 92]), AI-assisted decision-making (e.g., [143, 180, 189]), and cognitive reasoning (e.g., [175]). This scholarship relies on empirical work to observe human-centered factors and designs interventions to support better tool use. Experimental work has explained the influence of user mental models [12, 93, 99] and cognitive facets (e.g., biases, bounded rationality, critical thinking [20, 53, 92, 140]). Design interventions have shown the impact of various explanation types (e.g., global vs. local, example-based, counterfactuals [21, 51, 117, 127]) and interface designs (e.g., highlights, text percentages, cognitive forcing functions [18, 20, 97, 133]). As is evident from this rich body of work, interpretability tools have the potential to make data science more accessible, but at the risk of exacerbating inappropriate reliance on ML outputs. Many design interventions seek to reduce this inappropriate reliance.

Complementary to the individual cognitive and design facets described above is the growing knowledge that socio-organizational factors (e.g., job roles, team dynamics, hierarchies, policies, etc.) are equally important in considerations of appropriate reliance on interpretability tools. **We study one specific facet of hierarchical expertise development: how proficiency, the development of skill over time as people evolve from novices to experts, shapes the use of interpretability tools among data scientists.** While domain expertise has been the most commonly studied socio-organizational factor [28, 164], proficiency-based expertise can also represent a set of diverse perspectives, potentially affording another type of social redundancy against inappropriate reliance on ML. Indeed, HCI literature on distributed cognition [81, 86] and socio-organizational sensemaking [2, 91, 183] often cite this expertise development as a means of adding necessary human redundancy in effective task completion. Therefore, we ask the following research questions to understand the role of proficiency-based expertise in interpretability and data science:

RQ1. How do experts and novices approach data science tasks and tooling?
RQ2. How do experts and novices integrate interpretability into their workflows and build trust in their outputs?
RQ3. What do differences in expert and novice behaviors reveal about the nature and development of proficiency-based expertise in data science and interpretability?

To answer these research questions, we conducted contextual inquiries and semi-structured interviews with expert (N=12) and novice (N=11) data scientists to understand how proficiency-based expertise differentials manifest in data science tasks and tooling. Participants' data science expertise was established using a validated, objective ML literacy scale [84]; as well as qualitative signals of formal training and professional experience in ML. The study included pre-interviews about people's data science workflows; a contextual inquiry where participants performed an exploratory data science task set up in a Google Colab notebook; a set of multiple-choice questions about the data and model; and a follow-up interview on experiences and comparisons with their existing workflows. Our Colab setup included access to explanations from a popular tool, SHAP [118], as the main interpretability component of the study.

Our results show that proficiency-based expertise differentials significantly shape people's behavioral and subjective experiences with data science and interpretability. Novices and experts approach the data science task with completely different mindsets, with experts being driven by the purpose of the data and forming scoping hypotheses from the get-go. While novices in our study gravitated toward SHAP explanations as guiding narratives for exploration, experts integrated them selectively within these hypothesis-driven workflows. Their complementary orientations brought strengths and risks: novices' openness encouraged flexible learning but risked over-reliance, whereas experts' caution ensured critical evaluation but led to rigidity in task workflows. Taken together, we discuss the implications of these differences in mixing proficiency levels to create a productive balance in collaborative settings, and how interpretability tools should be designed to support both curiosity and critique.

[1]For brevity, we refer to these as interpretability tools going forward.

2 Related Work

2.1 ML Interpretability

Interpretable ML has become central to building trustworthy, transparent models for high-stakes use [115, 120, 167, 176]. Early work in the field focused on laying the theoretical foundations and establishing definitions [66, 98]. Doshi-Velez and Kim [60] defined interpretability as "the ability to explain or to present in understandable terms to a human," though the specifics of what that means and how to evaluate it remain elusive. Taxonomies like [14, 62, 132] provide categorizations of techniques; and surveys such as [152] formalize the objectives and highlight the technical hurdles that need attention. More recent surveys extend this work by offering different systematizations: [110] maps out a broad taxonomy of methods whereas [29] emphasizes metrics and societal impact.

In practice, interpretability is pursued either through glass-box models (e.g., decision trees, generalized additive models [95, 103, 107]) or post-hoc explainers for black-box models (e.g., LIME, SHAP, Integrated Gradients, SmoothGrad [116, 118, 147, 149, 168]). Recent directions extend interpretability into model design (causal structure, decoupled explanation, human concepts [88, 130, 142]) and examine how explanation types (global vs. local, example-based, counterfactuals) shape user understanding and behavior [21, 51, 117, 127]. Related desiderata—fairness, robustness, and generalization—interact with interpretability and influence real-world deployment [87, 89, 179].

2.2 Human Interaction with Explanations

Explanations often fall short of their intended goals and frequently lead to over-reliance instead [13, 20, 64, 93, 97]. Prior work attributes this to cognitive, design, social, and domain expertise-related factors.

Cognitive Factors. Explanations do not automatically improve comprehension and can overload users or draw attention to the wrong cues. For example, greater transparency sometimes reduces people's ability to detect errors due to cognitive burden [143], while logic-based or contextual signals can confuse rather than clarify explanatory information [93, 101]. Even the presence or absence of explanations shapes cognition and task perceptions differently [71]. This shows that interpretability is bounded by human cognitive limits [92, 140, 144].

Design and Contextual Factors. Explanation effectiveness also depends on design choices and task context. Explanations paired with feedback improve satisfaction, whereas explanations alone can frustrate [165]. The framing of information matters: explaining why a system acted can build trust, while explaining why it did not act can confuse [109]; presenting these details as questions vs. answers also matters [53, 127]. Too much detail can overwhelm users [96], and in low-resource settings, explanations can exacerbate over-trust [141]. Comparative and example-based explanations are usually more intuitive but can also backfire [23]. Interactive approaches can improve understanding but do not necessarily improve trust calibration [40, 92]. More recent work shows that over-reliance depends on task difficulty, explanation clarity, and incentives, with well-designed explanations lowering the cost of verification and improving calibration [175].

Social Factors. Explanations also shape social judgments of AI competence, expertise, and authority. They can increase acceptance of AI decisions without improving calibration [13, 20, 102, 189]. People view AI as more competent in technical domains than moral ones [173], and sometimes perceive AI as fairer than humans [8, 78]. Explanations interact with advice-taking: users lean on algorithms unless they have strong confidence [30, 113], but abandon them quickly after mistakes [59]. Group-level work shows similar dynamics: explanations can increase reliance on AI, but diversity of perspectives helps catch errors [44].

Expertise Differentials. Reliance also differs by domain expertise. People evaluate explanations against disciplinary norms [4, 138, 170], and explanations are more effective when users have relevant background knowledge [180]. Users' cognitive constraints and biases shape how they interpret and rely on explanations [1, 16, 70,

91, 140, 156], and interpretability practices vary by organizational role [83]. Explanations often help only those unfamiliar with a task [154], and tailoring details to expertise improves outcomes [72]. Novices benefit when barriers are lowered through dialogue, targeted support, or AI-mediated creativity tools [45, 55, 56, 187].

Our work builds on human-centered scholarship showing that the accountability promise of interpretability comes with contextual challenges. Explanations can amplify rather than reduce bias [78], fail to empower non-experts even when designed to [17, 102], depend on how users make sense of them [91, 127, 156], and are sensitive to technical choices invisible to most users [87]. We extend this by examining how proficiency-based expertise shapes reliance on interpretability tools, identifying for whom explanations work and under what conditions they support accountability.

3 Methods

3.1 Main Study Flow

To understand how proficiency-based expertise differential impacts data science and interpretability tool use, we conducted a study with data scientists (N=23), comprising three components: (1) a context-building pre-study interview; (2) a contextual inquiry using a think-aloud protocol, where participants conducted exploratory data analysis using SHAP as the interpretability tool; and (3) a semi-structured follow-up interview. We used SHAP in our study given its widespread adoption in both research and practice, its open access, and its consistency with other feature-attribution methods such as LIME [147], EBMs [139], and the What-if Tool [185].

As a first step, we recruited participants using an intake survey to introduce the study, get informed consent, verify participant eligibility, and establish their background and expertise in ML and interpretability (see Appendix A). To effectively manage recruitment based on proficiency, we relied on Hornberger et al. [84]'s validated objective AI Literacy Scale: an objective scaled allowed for easy scoring of proficiency, and a validated scale added credibility; this scale was the only one available at the time that met both criteria. From this intake, we reached out to a subset of survey responders to take part in the main study based on their results. We briefly overview our study artifacts, including the intake survey, dataset, SHAP tutorial, model and explanation outputs, and MCQs; more details on specific questions and protocol can be found in Appendices B and C. All study procedures and protocols were reviewed and approved by the University of Minnesota Institutional Review Board.

3.1.1 Pre-study Interview. We first asked participants to keep their video on throughout the session to ensure legitimate participation. We confirmed their ML knowledge and experience through some open-response questions, and asked them to describe their typical ML workflow, covering data selection, pre-processing, model selection and validation, and deployment. These workflow questions were asked both generally and using their most recent data science experience as a retrospective anchor [85] to confirm details.

3.1.2 Contextual Inquiry. We conducted a contextual inquiry that followed a think-aloud protocol, using a shared Google Colab notebook containing the study task. The Colab notebook was organized into sections by cells: imports, dataset exploration, model training, visualizations from SHAP-based explanations (see Appendix Figure 2 for example SHAP plots from our Colab notebook), and post-exploration MCQs. The task was based on the Titanic dataset[2] which is publicly accessible and does not require esoteric knowledge. Participants were asked to conduct exploratory data analysis on the included dataset while referring to the provided SHAP tutorial (Appendix C.2). They were encouraged to write their own code and run independent analysis, and use SHAP to support their exploration as appropriate. Throughout this part, following think-aloud protocol norms [136], participants were asked to verbalize their thinking and were periodically nudged by the researchers to do so.

[2]https://www.kaggle.com/competitions/titanic/

3.1.3 Post-Exploration Multiple Choice Questions (MCQs). After exploration, participants moved to the final cell of the Colab notebook, which contained 10 MCQs to assess their reasoning skills with model outputs using global and local explanation interpretation, drawing on prior interpretability work [92, 93, 108] (see Appendix C.5 for details). Participants could answer using SHAP or their preferred data science methods, and rated their confidence for each answer on a 7-point Likert scale.

3.1.4 Semi-Structured Follow-up. Finally, we conducted a semi-structured interview to understand participants' reliance patterns and attitudes towards interpretability tools. We asked them to reflect on whether they wanted to rely more on SHAP or their own intuition, along with a justification for their (lack of) reliance. We also inquired about instances where they blindly trusted a SHAP plot and their overall opinions on the visualizations. Additionally, they were asked to assess their self-perceived expertise in ML and qualitatively evaluate their own task performance.

3.2 Participants and Data

Participants were first recruited via our intake survey shared on ML-oriented subreddits and LinkedIn. The intake survey was a critical aspect of our methodology – establishing a grounded way to determine proficiency shapes our research questions and results. We received 400 responses with high scores on the objective AI literacy scale. However, it became clear in interviews that most of these participants' ML and AI knowledge did not match their scores, and they were using generative AI during the main study. Therefore, we revised our recruitment strategy, changing our outreach from public forums to university newsletters, industry-specific Slack channels and LinkedIn groups, and snowball sampling through trusted networks.

To establish participants' status as novices or experts, we operationalized proficiency using two approaches. First, we used the validated AI Literacy scale [84], grouping participants based on defined score thresholds: novices scoring between 33% and 66% (inclusive) and experts scoring 67% or higher. We chose these cutoffs based on the distribution of scores and selecting a split that both reflected meaningful differences in task performance and yielded comparably sized groups for analysis.

Second, we confirmed participants' proficiency qualitatively during pre-study interviews, where we asked them to describe their ML experience, typical workflows, and decision-making processes, recognizing that expertise is situated in professional practice and not fully captured by standardized measures alone [19, 46]. In rare cases where a participant's demonstrated knowledge in the interview diverged from their literacy score, we adjusted categorization based on their described experience prior to task engagement. In the first recruitment round that we later disregarded, we had also included a *subjective AI literacy scale*: MAILS (Meta AI Literacy Scale) [27]. Almost all participants rated themselves highly, making it ineffective for distinguishing expertise levels. We removed the scale from the main recruitment efforts and disregarded its results.

Our revised recruitment call received 290 responses. Based on defined thresholds, 108 experts and novices qualified, and 23 completed the study: 12 experts and 11 novices. Further recruitment was stopped after all authors agreed that our data trends had reached saturation and our sample size (N=23) was sufficient as per guidance on grounded theory qualitative methods [24, 48]. On average, sessions took 67.6 minutes, and participants received $40 compensation. Participants were aged 21–30 (M=25.3, SD=2.7), most held graduate degrees (16 MS, 5 PhD, 2 BS), and 16 identified as male and 7 as female. Mean self-reported ML knowledge was 5.0/7 (experts: 5.4; novices: 4.5) and interpretability-tool familiarity was 3.6/7 (experts: 4.3; novices: 2.7), with objective literacy scores ranging from 38.7% to 93.6%. Individual participant demographics are provided in Appendix Table 1.

3.3 Analysis

We used Zoom-generated transcripts of the study, with manual edits to fix any inconsistencies. All verbal data—including the contextual inquiry and interview components—was transcribed and analyzed holistically rather than by study phase. Participants were assigned anonymous identifiers, and no identifying information was

recorded in the transcripts. Our analysis was based on a grounded theory approach [31, 32, 48]. In the first stage, three researchers used MaxQDA[3] to open-code all 23 transcripts independently. They met regularly to discuss progress and note emerging observations. Because our research questions were exploratory, the coding process was iterative. Codes served as tools for developing emerging themes through constant comparison, not fixed labels to be tested. Accordingly, we did not calculate inter-rater reliability [122]. After finishing open coding, participant categories—expert and novice—were reintroduced. In the second stage, axial codes were formed using affinity diagramming (see process in Appendix Table 3). These axial codes were compared across participant categories, and thematic clusters were developed to address the research questions. To get diverse perspectives, an additional researcher not involved with open coding participated in the second stage.

4 Results

4.1 Descriptive Statistics and Overview

Novices performed better than experts on the 10 post-task MCQs (accuracy for novices: M = 80.9%, SD = 16.4; experts: M = 74.2%, SD = 17.3). Self-reported confidence on the MCQ responses, measured on a 1–7 Likert scale, was similar (novices M = 5.95, SD = 0.42; experts M = 6.1, SD = 0.47). These numbers are a result of different analytic styles more so than ability. Novices inspected SHAP plots comprehensively, which improved their recall, while experts skimmed selectively to integrate explanations across the task (details in subsections below).

We also report time as a percentage of each participant's total interview time. Experts and novices distributed their time similarly overall, with some meaningful contrasts. Experts spent slightly more time on dataset review (M = 3.6%, SD = 1.9) and model exploration (M = 2.7%, SD = 1.3) than novices (M = 2.7%, SD = 1.6; M = 2.2%, SD = 1.0; respectively), indicating a stronger emphasis on orienting to the data and modeling setup. Both groups devoted about one-third of their time to SHAP exploration (experts M = 32.1%, SD = 6.1; novices M = 31.6%, SD = 6.2). Novices, however, allocated substantially more time to the SHAP tutorial (M = 12.1%, SD = 5.0) compared to experts (M = 8.7%, SD = 2.0). This comes up again in our qualitative findings, with experts relying more on their priors and novices leaning more heavily on SHAP.

4.2 Approaches to Data Science Tasks and Tools

4.2.1 Critical vs. Exploratory Approach. **Experts** approached the task with a more analytical mindset; they engaged *critically*—questioning the study setup, evaluating its components, and considering its broader purpose and implications. Experts asked for justifications of our workflow choices, for example, "[I want] a lot more data validation before modeling" (P14); "you didn't do PC analysis at all...for feature engineering?"; and "trying alternatives like LightGBM... better for computational efficiency issues" (P15). Their think-aloud comments reflected comparisons between our setup and their own workflows. Compared to experts' problem-definition driven critique, **novices** did not define a problem going into the task. Rather, they adopted an *exploratory* approach, developing their understanding incrementally as they progressed through steps linearly. When asked about their approach beyond the study, they mentioned standard data science practices like checking data distributions or missing values, but without providing a reasoning. This was akin to reading from a textbook, wherein they described their process as general habits learned in class rather than as actions tied to specific goals. As P4 described:

> "It's first going to be pre-processing, right? Identifying [how many] net total features are there? What is my dataset? What is my sample size? Am I going to use pandas or am I going to use PySpark?" (P4, Novice)

4.2.2 Data-First vs. Prediction-First. The differences in expert vs. novice behaviors in completing the task were noticeable from the get-go with how they each approached the role of the dataset. For **experts**, data science was

[3]https://www.maxqda.com

all about the data. They preferred a top-down approach where they "first understand the objective when I have a new dataset" (P19); formed hypotheses about the prediction task; and then asked questions, probed assumptions, and evaluated the workflow in the context of their data understanding. This sometimes included looking for "prior work that works [i.e., develops models] in such settings" (P15) or "reaching out to domain experts" (P20) if the dataset was in an unfamiliar domain (e.g., medicine, finance).

Once they understood the broader context of the dataset, they moved on to evaluating the specifics. While experts conducted the standard quality checks (e.g., managing missing values and outliers, analyzing data distributions), they also experimented with the data, removing specific data points to observe how patterns shift, merging correlated features, or eliminating irrelevant variables to refine their understanding of the data structures. They formed hypotheses about trends and then evaluated them by applying, for example, visualization techniques. P20 shared:

> "While pre-processing, I first visualize and then use these visualizations to inform pre-processing because I sometimes have initial hypotheses. For example, whether cabin is related to pclass and can it be [meaningfully] used?" (P20, Expert)

Novices, by contrast, took a more bottom-up approach, relying on model predictions and SHAP local explanations to guide their understanding of the data and task. They quickly moved from the data to the model and SHAP outputs. Many, like P12, made quick model choices without much data exploration:

> "If we have a lot of number of records, it's a more complicated dataset then I would lean towards deep learning. If it's less [data] or seems like a naive machine learning algorithm can deal with it, then probably go with that." (P12, Novice)

When they got to SHAP, novices did not verify the data using alternative techniques like experts. Their evaluation relied on reviewing individual cases and reasoning about prediction accuracy using SHAP.

> "So I remember it [SHAP dependence plot] was saying earlier that if you were male you would have a higher likelihood of surviving versus female so I think this case [individual prediction] seems to align with what I saw earlier [dependence plot]." (P6, Novice)

4.3 Integrating Interpretability Tools into Data Science Workflows

Our findings show that SHAP enters expert and novice workflows in different ways: experts use SHAP selectively—as a check that integrates into an existing hypothesis-driven workflow—while novices use SHAP systematically—as the anchor for exploration, evaluation, and sensemaking. Here, we first examine SHAP as a means of narrative sensemaking, then contrast selective attention vs. systematic inspection of SHAP visuals, and finally show how SHAP is more deeply embedded in exploratory rather than critical workflows, shaping trust and reliance differently across proficiency levels.

4.3.1 SHAP as a Tool for Narrative Sensemaking. Human sensemaking is inherently driven by narratives [91, 114, 183], and we find that SHAP can be a helpful tool if used appropriately for this purpose. While both experts and novices craft narratives to make sense of the task, experts rely on their intuition and novices gravitate towards SHAP outputs as a guiding narrative.

Novices use SHAP visualizations as narrative anchors that shape their understanding of the model performance and data. The explanations make it easy for novices to generate narratives for why a prediction was made and hypotheses for trends they want to explore on subsets of data. For example, during data exploration and answering MCQs, novices note:

> "Kind of obvious that the class with the highest fair, like business class, has the most survival because probably they are in more secure environments...have like more boats for emergencies." (P12, Novice)

> "As age increased, that's probably a factor of why they didn't make onto the lifeboat. And this would make me curious about middle-aged or people who have children, how their likelihood was different from people who didn't have children." (P17, Novice)

SHAP enables this kind of narrative sensemaking, both at the global level (e.g., seeing different values of a feature and its impact on the outcome – dependence plots) and local level (e.g., seeing how different input feature values result in a given outcome for an individual – local bar plots). This can yield positive outcomes compared to intuition-driven sensemaking: many novices set aside their own intuition and tried to craft plausible narratives from the SHAP visuals. Indeed, this helps them perform just as well as **experts**, who can sometimes overfit to their own opinions or intuitions about the task and model performance. However, appropriate reliance in both cases depends on the accuracy of underlying assumptions: for novices, the assumptions SHAP makes about the model; for experts, the assumptions behind their intuition.

4.3.2 Selective Attention vs. Systematic Inspection of SHAP Visuals.

Experts demonstrate selective attention during data analysis, focusing on relevant components. They go through the notebook intentionally—skimming through "obvious code blocks" (P20) and identifying which parts of data, model, and SHAP visualizations need further evaluation. For SHAP, experts adopt a strategic approach: they read the tutorial, examine one representative plot, and then quickly skim through similar visualizations.

> "So these are the things we can take away. I guess I might not remember all of this [SHAP plots], but I will come back to these if needed." (P20, Expert)

They attribute this selective behavior to experience, "familiarity with tools like SHAP" (P20), "similar charts and outputs" (P2), and common data science setups. They pay more attention to elements that help them decipher the task—identifying data types, "whether image, spatial, or time-series" (P3), determining problem formulation (regression versus classification), and scoping its application. This allows them to build a mental framework first, which they later supplement with details from SHAP if needed for the MCQs.

Novices have a similar focus on details, but they gravitate towards the details of outputs rather than code. They are systematic in their evaluation of SHAP. Novices read SHAP plots closely, ask more questions than experts—both during the tutorial and in follow-ups—and hypothesize what each visual is trying to communicate. They memorize details of visuals which they reiterate later when answering MCQs. When interpreting visualizations, their questions focus on visual elements: "color meanings" (P16), "axis ranges, legend interpretation, visual marks" (P4), and underlying representations being used (e.g., "absolute SHAP values or not" (P10), "purple in decision plots...same as dependence plots?" (P13)). They treat SHAP values as valid model approximates.

4.3.3 SHAP is More Extensively Integrated into Exploratory Rather Than Critical Workflows.

Most participants had some experience with interpretability tools, and many (especially experts) routinely used similar tools. While SHAP could be easily integrated into workflows of both, novices show more eagerness to integrate it.

Experts' workflows are not entirely reshaped by SHAP; rather, it plays one small role in their original workflow. They maintain their analytical approach, integrating SHAP as a support tool to validate pre-existing hypotheses. They do not base their analysis or conclusions on SHAP. Moreover, experts ask technical questions to understand the underlying assumptions of SHAP, specific visualization types, "how SHAP values are calculated" (P12), hyperparameter effects and methods for synthesizing information across multiple plots. These details are intended to help them calibrate reliance on SHAP. Throughout this integration, experts maintain a diverse toolkit approach that preserves their workflow while giving the "new" interpretability tool a supportive role.

Novices' workflows are more fundamentally influenced by SHAP: it becomes the center of exploration and evaluation, and the baseline they compare everything against. As novices do not start their exploration with task-driven hypotheses (see Section 4.2.1), they adapt their exploration to what SHAP can reveal, prioritizing visualization-driven insights.

> "It [SHAP] was really helpful to understand patterns. If I had to do this manually by plotting graphs for everything, seeing correlation and everything, it would be more time consuming. But this [SHAP] was pretty easy to understand once you get the hang of it." (P13, Novice)

4.4 Trust and Reliance on Interpretability Tools

4.4.1 Experts' Skepticism and Under-Utilization of SHAP. Experts engaged with SHAP through a lens of conditional trust rather than as a definitive source of data or model understanding. They cross-checked SHAP explanations against their domain knowledge, and often dismissed outputs. Their skepticism arose from three main reasons.

First, experts described a paradox where a tool designed to increase transparency actually introduced a layer of ambiguity. Despite reading the tutorial and doing their research on it, they could not decipher "how SHAP values are calculated" (P21). Given this, they preferred standard data science methods. As P19 pointed out:

> "What does it [SHAP values] actually represent in terms of predicting...what's the background calculation. I'm still not sure about that so if we have that calculation available and if I put these values in a linear equation to tell whether these are slopes then it can be more interpretable and I can do better in terms of decision making but right now I'm not sure if minus 0.86 is actually a slope." (P19, Expert)

In most cases, this resulted in experts' ignoring SHAP, instead writing and running their own code to confirm assumptions or testing alternative modeling or interpretability approaches that they considered to be "more appropriate analysis" (P15), for example, relying on a "pure mathematical model instead" (P2).

Second, experts found SHAP's visual presentation of information unnecessarily complex. When going over individual plots, they cited other ways in which the same information could be presented:

> "Instead of looking at SHAP value... I would just plot the age distribution... These explainability tools are great for certain tasks but sometimes the question can be answered more directly by something simpler or even something more complex but not necessarily using visual tools." (P15, Expert)

While critiques like these were common from experts, they also ultimately preferred these more complex visuals. Experts particularly critiqued complex plots showing partial dependence and decision subsets, or dense tables of counterfactuals and semi-factuals, citing these to be counterintuitive. However, they ultimately preferred these visuals over global and local bar plots, calling the latter "too simplistic" (P20) and "potentially biased" (P19). We describe these preferences in further detail below (Section 4.4.3).

Finally, several experts mistakenly attributed data and model issues to SHAP, amplifying their skepticism in its validity. Our participants noted some outputs with potential biases from overfitting to anomalous data points. However, instead of associating these with the data and modeling aspects of the pipeline, they used it as further evidence to under-utilize SHAP. A rarer subset of experts (2 in our sample) correctly delineated SHAP from the data and model issues, and acknowledged the opportunity that SHAP visuals afforded in identifying the potential data and model biases noted here.

4.4.2 Novices' Enthusiasm and Over-Utilization of SHAP. Novices treated SHAP as the sole interpretive anchor and ground truth, placing confidence in its outputs since they did not have the experience of other established ways of discovering similar, interpretable information. They referenced SHAP plots as being more "accurate than their own intuition" (P4) and "providing novel information that is useful" (P8). Their understanding of the entire setup often resulted solely from SHAP outputs. They referenced SHAP plots alone while explaining their reasoning behind MCQs and did not deliberate further:

> "Answer is [option] 4 based on the plot. It shows that females were more likely to survive." (P16, Novice)

Novices' enthusiasm about SHAP's visual approach to interpretability is exactly what the tool is designed for, but results in over-utilization due to novelty bias and lack of expertise. Novices might lack prior experience to inspect

the data and model in ways that improve their understanding. In this scenario, SHAP is an ideal alternative: visual representation of otherwise more complex information. Indeed, SHAP, like all visual interfaces, capitalizes on visual literacy in communicating complex data information. Novices quickly learned to interpret these visuals compared to the equivalent learning of data and model. We hypothesize that this is why novices over-rely on SHAP.

4.4.3 Differential SHAP Preferences. Expert and novice behaviors were consistently different in both their trust and reliance judgments, and their preferences for SHAP visuals. While experts preferred the more information-rich charts, novices prioritized ease of interpretation and learning with the simpler bar plots.

Experts described the denser SHAP outputs as "particularly intriguing" (P15) because they allowed them to make objective observations on their own. They relied primarily on the counterfactual tables, decision plots, and partial dependence plots, all of which the novices explicitly mentioned disliking. As some experts noted:

> "Can I rank them? I feel that most of them were helpful but PDPs were the most helpful." (P11, Expert)

> "Counterfactual table was much more useful as far as I know because of the comparison thing. With this one table, we can manually interpret whether subsets of people survived or not. And why. For other plots, you need [to write] code to gather more information for reasoning. But with this, you can more easily brainstorm and do rough hypothesis building. (P21, Expert)

For experts, plots with clusters of data or relationships between multiple variables offered opportunities to compare data points, see patterns, and draw their own conclusions. They were skeptical of simple outputs and did not want to overfit to the narratives salient in the simpler plots; they wanted the room to make independent, seemingly objective observations. In their view, the real value of the tool came from the complex plots that enabled deeper analysis.

An interesting contradiction emerged: while experts preferred these complex plots, they were also frustrated by the lack of transparency in how the outputs were calculated and about certain presentation decisions. They wanted visual representations of the SHAP calculation process to appropriately evaluate the explanations:

> "I'm still not sure how SHAP in the background for XGBoost is calculating these values for the chart axes. It's aggregating to tell that class importance is 0.6 for a specific age in the PDP. But I'm not sure how it's quantifying this against other leaf [feature] nodes. So it depends on how SHAP created the TreeExplainer, what are the settings for the tree? If you are using too many tree nodes you could of course overfit the data as well. Then slight changes can change the explanation entirely. I can't confidently take decisions here because the selection of features and ordering for plots depends on these choices." (P19, Expert)

Novices were not inherently skeptical of SHAP and preferred simpler, easy-to-interpret outputs free of complex visual marks and channels (e.g., local and global bar plots). Because they trusted SHAP outputs, they wanted plots that afforded quicker decision-making with minimal information to process, which is exactly the opposite of experts' approach. They thought of complex plots like PDPs and counterfactuals as "cluttered" (P4), "hazy" (P13), "confusing at first" (P6), and "noisy" (P23). Early in the study, they tended to misread these plots, though their accuracy gradually improved.

4.5 Experience-based Maturity

Beyond static differences, we observed participants' approaches evolving during the study. We call it *experience-based maturity*: the ways prior knowledge and openness to new input shape how people adapt their strategies in real time. In our context, this appeared in how participants decided when to rely on SHAP versus intuition, how often they revisited materials, and in the language they used to describe their workflows.

4.5.1 Transfer Learning vs. Flexible Learning. **Experts** displayed transfer learning driven by efficiency. They drew on established schemas and workflows, connecting new information to familiar mental models. This afforded

efficiency, but also rigidity. After initial exploration, many experts felt they had "enough" understanding and stopped consulting specific SHAP plots, leaning instead on intuition shaped by early impressions. P20 noted:

> "I would want the data to inform my understanding of what went wrong so once I look at explanations I try to justify it against my intuition...I would call it more reasoning than intuition." (P20, Expert)

This approach reduced effort but risked over-reliance on initial hypotheses and under-use of SHAP's insights.

Novices, by contrast, demonstrated flexible learning driven by curiosity. Their workflows were not rigid, allowing for the use of SHAP in its intended use-cases. They relied heavily on SHAP, exploring its outputs with curiosity and an openness to surprising insights. They referred to the same plots multiple times. As a novice said:

> "I still relied more on plots because I would go back to confirm before answering anything. I think exploring the data this way showed me some surprising results, which is why I wasn't totally relying on intuition because there was some kind of seemingly contradictory or surprising results." (P17, Novice)

This flexibility surfaced edge cases that experts overlooked, while introducing risks when outputs were misleading.

4.5.2 Different Terminology. These learning styles were also reflected in language. Experts used terminology grounded in practice, drawing on lived experience rather than textbook phrasing; their language layered personal judgment onto formal knowledge. Novices, by contrast, relied more on textbook language, repeating codified techniques learned in coursework. Their language lacked the nuance that comes from applying concepts across different situations. This shift from formal, decontextualized terminology towards nuanced terminology captures one way expertise matures and is an important signal of that deliberate practice of gaining proficiency.

5 Discussion

To situate our contributions in the broader expertise literature and provide a theoretical lens for interpreting our findings, we first outline a framework of expertise grounded in a scoping review of this prior work [9]. This framework helps explain the different cognitive, organizational, and runtime dimensions of proficiency-based expertise development. We provide a summary of our results and their connection to the Expertise Framework in Figure 1.

5.1 A Framework of Expertise

In developing this framework, we started with classic studies of expertise and used forward citation snowballing to trace how these ideas are applied across domains. Expertise development has been a topic of research across many fields since the 1940s, with early research centered on cognition in chess and card games in lab studies, identifying characteristics that distinguish experts from novices [37, 57, 69]. Building on this, field experiments studied the differences in expertise in various topical domains [3, 49, 104, 123, 184]. From these studies, it became widely recognized that expertise grows over time rather than being innate [6, 182]. Additionally, linear, stage-based models of expertise development were critiqued, offering an alternative: that becoming an expert involves a transformation in how individuals perceive and engage with their work, not just an accumulation of knowledge or skills [52]. We synthesize the cognitive processes, knowledge development and organization, and runtime task strategies relevant to this transformation into expertise (summarized in Appendix Table 2).

5.1.1 Cognitive Processes. Cognitive processes refer to the procedural aspects of cognition involved in acquiring, representing, and using knowledge. These processes include chunking, pattern matching, system thinking, and elements of memory and recall.

Chunking. Experts differ from novices in how they mentally organize information, forming larger and more complex cognitive chunks that streamline their workflows [63, 162]. This includes experts recognizing meaningful chunks of information that novices see as isolated elements [57], and indexing more complex chunks as one unit instead of smaller pieces of information at a time [38]. This has been verified with field studies across domains [49, 57].

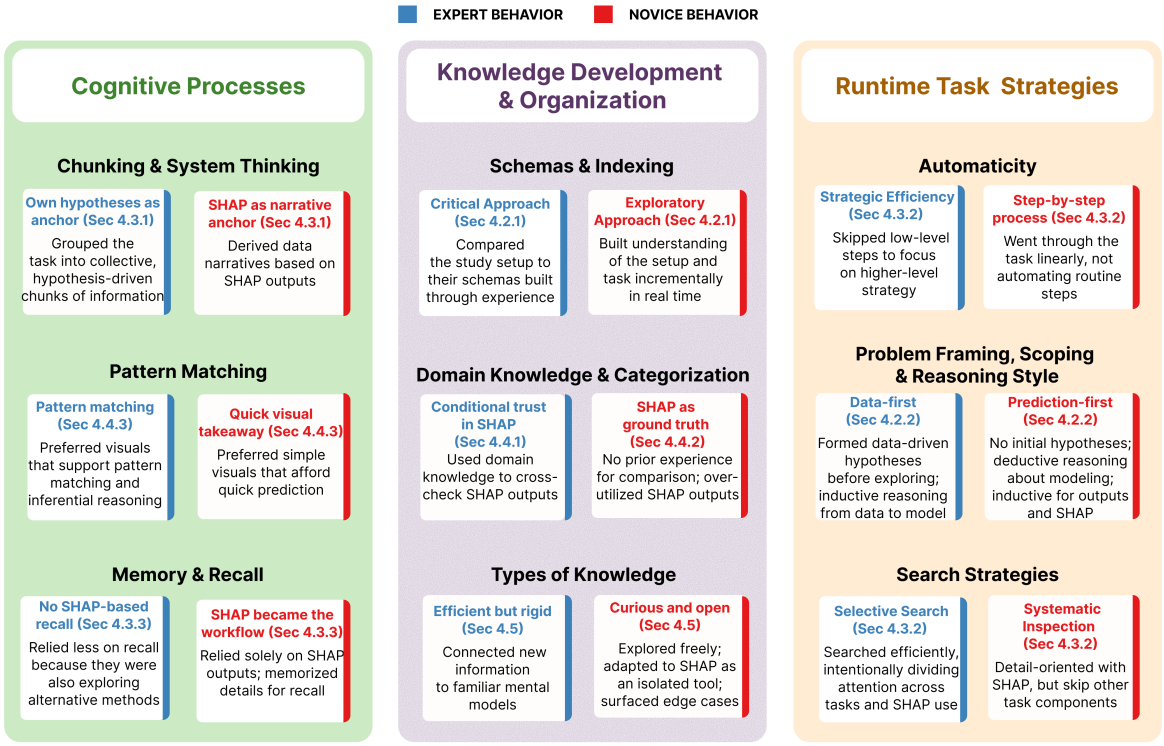


Fig. 1. Summary of qualitative findings contrasting expert and novice behaviors, organized by the dimensions of our expertise framework. Each behavior pattern is linked to the corresponding results section. The framework spans three high-level dimensions: Cognitive Processes, Knowledge Development and Organization, and Runtime Task Strategies, showing how expertise shaped participants' behavior regarding data science practices and SHAP explanations across each dimension.

Pattern Matching. Experts recognize patterns and relevant cues faster, drawing on well-organized knowledge chunks. Novices, on the other hand, lack the ability to identify shared cues across information types that can enable faster prediction of outcomes [100]. This was initially studied in chess, [25], and verified in other domains [77, 121, 123, 171].

System Thinking. Experts adopt a systems perspective, reasoning about how components interact to produce higher-level behavior [15, 150]. Novices focus on elements in isolation, while experts interpret problems holistically [58, 119].

Memory and Recall. Work in cognition often cites the capacity of one's short-term memory as the constraint on human ability to think, solve problems, and process information [11, 125, 135]. This would put experts and novices on an even footing. However, experts' superior recall ability cannot be explained by short-term memory capacity [35, 36, 76, 146]. Classic work with chess players demonstrated that experts recall meaningful positions better than novices, but lose this advantage when positions are randomized, pointing to their dependence on organized long-term memory [33, 41, 57].

5.1.2 Knowledge Development and Organization. Expertise also depends on the development of mental schemas that help with recognizing patterns, categorizing concepts, and different types of knowledge.

Schemas and Indexing. Experts draw on interconnected schemas built through experience, which help them identify underlying structures [26, 43, 172] and filter out irrelevant information, but it can also create rigidity—experts may overlook novel insights that fall outside their established frameworks. Novices, by contrast, approach problems more openly, without preconceptions. While this makes it harder for them to filter information or prioritize effectively [100], it allows them to consider more possibilities that experts might dismiss.

Domain Knowledge and Categorization. Experts and novices also differ in how they categorize and approach problems. Because of deliberate practice, experts spend more time interpreting a problem but are more efficient overall [6]. Rather than memorizing solutions, experts draw on domain experience to categorize problems in principled ways, ignoring surface details that novices often over-focus on [49, 184].

Types of Knowledge. Building on prior work [50, 68, 74], Suresh et al. [169] distinguish between three types of knowledge that delineate expertise: formal, instrumental, and personal. Formal knowledge refers to the theoretical and factual information gained through formal education; this is often consistent across both groups [42, 178]. Instrumental knowledge captures how formal knowledge is applied in real-world contexts (e.g., experience with specific tools or tasks); this category develops over time with experts having an advantage [74, 105]. Finally, personal knowledge involves the experiential and intuitive resources individuals bring to situations that are developed informally through observation and practice [65, 67]. In complex settings, decision-making often relies less on textbooks and more on this accumulated personal knowledge [47, 61]—this shows the need for deliberate practice in the development of proficiency.

5.1.3 Runtime Task Strategies. Runtime strategies describe how people reason during the task, how automatic their actions are, how they frame problems and how they search.

Automaticity. With practice, experts shift routine parts of tasks into automatic processes [73, 174, 186], allowing them to focus on strategy. This has been studied under various theoretical names (e.g., the dual process model [90], bounded rationality [161]). The patterns that experts develop through repeated exposure trigger an efficient and automatic course of action [112, 145, 148]. These automatic routines enable speed but can break down when conditions change [75]. Novices complete a task step-by-step, but can recall larger chunks of solutions automatically over time.

Problem Framing and Scoping. Experts scope problems early and iterate quickly [10, 157] navigating the tasks effectively. Novices spend longer on instructions before acting, reducing early errors but limiting progress [106].

Reasoning Style. Novices lean on deductive, rule-based reasoning, limiting flexibility, while experts are more inductive, beginning with tentative solutions and refining them through action [61, 106]. Experts' inductive approach makes them better at evaluating their own understanding and correcting errors, whereas novices rely more on trial-and-error and struggle to self-monitor [34, 80, 129, 134, 166].

Search Strategies. Experts differ from novices in both the effectiveness and efficiency of knowledge use. They conduct structured, knowledge-driven searches, whereas novices rely on undirected, exhaustive strategies [69]. Experience allows experts to recognize relevant information quickly [148] and process problems more rapidly [38]. Novices, by contrast, can often get stuck in information gathering before advancing to problem-solving [49].

5.2 How this framework applies to our work

Our study demonstrates that the dimensions of expertise outlined in our framework provide a foundation for anticipating and interpreting how novices and experts approach data science tasks and interpretability tools. We summarize the key connections here before turning to their implications for design. Similar to the framework's *cognitive processes*, experts in our study treated the data science task in collective chunks that were driven by

hypotheses about the data; meanwhile novices focused on isolated details of data and SHAP, memorizing short-term units. For *knowledge development and organization*, we noted experts' focus on instrumental and personal knowledge, and efficiency in building interconnected schemas between SHAP outputs and their data-driven hypotheses; novices began with formal knowledge alone, but adapted to SHAP more easily by indexing it as an isolated tool. Experts' domain knowledge also served as a check on SHAP. They could cross-reference outputs against what they already knew about the data, while novices, lacking this reference point, accepted SHAP outputs at face value. As expected with *runtime behaviors*, experts scoped the problem early and applied inductive reasoning for evaluation. They searched selectively, skimming familiar plots and focusing on higher-level strategy, while novices inspected each plot exhaustively. However, experts were rigid in following their intuition—novices' openness to SHAP enabled performance comparable to or better than experts.

The consistency between our findings and decades of expertise research expands two paths forward for interpretability: (1) designing to guide individual practitioners toward better tool use; and (2) leveraging proficiency-based differences as a form of social redundancy for appropriate reliance.

5.2.1 Improving Interpretability Tool Design for Individuals. Since the use of interpretability tools is filtered by expertise as per the framework above, we consider HCI design strategies that prioritize the distinct needs of experts and novices. Experts in our study approached SHAP selectively, skimming familiar outputs, writing their own code to verify results, and expressing frustration with the opacity of SHAP's underlying calculations. These behaviors point to two design needs. First, interpretability tools should provide *accelerators* for experts [137]—shortcuts, hidden from novice users, that speed up the explanatory interactions for experts. These can be interaction shortcuts for rapid navigation across plots, customizable workflows that match hypothesis-driven analysis patterns, or API access for programmatic integration of SHAP calculations into existing coding consoles. Second, experts' skepticism about SHAP's computational opacity calls for transparency on demand [160]: while all users may begin with global summaries, experts should be able to drill down into computational provenance (e.g., TreeExplainer parameters), sensitivity analyses showing how SHAP values change with different parameters, and side-by-side method comparisons that let them situate SHAP outputs against approaches they trust. This layered approach prevents overwhelming novices with complexity while giving experts the methodological transparency they demand to calibrate appropriate reliance.

Novices treated SHAP outputs as ground truth, a pattern rooted not in carelessness but in the absence of established workflows. This points to a different set of design heuristics that prioritize *error prevention* and *scaffolding* [137]. Interpretability tools could offer verification prompts that encourage cross-referencing SHAP claims with raw data; contextual tooltips with progressive disclosure that make more methodological details accessible over time; and structured exploration paths that guide users through global patterns, hypothesis formation, and contradiction-checking. The visibility of deeper complexity layers—even if novices do not always access them—can itself build appropriate skepticism by signaling that explanations involve assumptions and alternatives [91]. We observed signs of proficiency development within our study sessions: novices who initially called complex plots like PDPs "cluttered" became noticeably more accurate with them by the end after being made to think aloud during the study. This suggests that interpretability tools could support the transition between these proficiency levels, but require designs that offer scaffolding—reflective designs that prioritize critical thinking [159], narrative visualizations connecting different explanation types [158], and gamified designs that leverage curiosity [177] may be useful paths forward. Together, these design principles address the dual risks of under- and over-reliance that our framework predicts, while treating proficiency as something tools can help develop rather than just accommodate.

5.2.2 Distributed Cognition and Social Redundancy for Appropriate Reliance. We observed complementary strengths and risks across proficiency levels that could be leveraged for better collective work. Novices' exhaustive strategies surface alternative insights and edge cases critical to ML, while experts' efficient reasoning provides

necessary friction against tool over-use. Rather than steering novices toward expert-like behavior, interpretability tools should scaffold strategies that blend novice exploration with expert rigor. Since our study involved individual sessions, we did not observe direct expert-novice collaboration, but our findings suggest that design directions for collective work are worth exploring. Indeed, this is a form of social redundancy that HCI research on distributed cognition and organizational reliability suggests makes mixed teams more reliable [86, 183].

To enable this, for example, tools could support shared artifacts where team members record their interpretations independently before converging so that novices' exploratory approach is not prematurely influenced by experts' tendency to discount SHAP in favor of other methods. In our study, novices and experts often reached different conclusions from the same plot. Disagreement logs that show divergent interpretations for collective review could provide a space for these differences to be discussed. In settings like model audits, role-differentiated views could present the same explanation at different levels of complexity, ensuring that insights from both proficiency levels are preserved rather than one dominating the review.

6 Limitations and Future Work

We acknowledge several limitations for external validity. First, our contextual inquiry took place in a Google Colab notebook, which may not fully reflect the open-ended nature of real-world data science work; future studies should embed interpretability tasks in more situated workflows. Second, our focus on SHAP constrains the generalizability of our findings to this method; future research should test whether similar expert–novice differences arise with other explanation methods and tools. Finally, because our study was short in duration, we observed only initial signs of learning and adaptation. Longitudinal observations may offer more detailed insights into expertise.

7 Conclusion

Interpretability tools are designed to increase accountability and trust in machine learning, yet their efficacy in practice is complicated by several human-centered factors. Our study describes the impact of one such factor: a proficiency-based expertise differential (i.e., novice vs. expert proficiency), which fundamentally alters how interpretability tools are used. Via a contextual inquiry with expert and novice data scientists (N=23), we find that novices bring flexible curiosity but risk over-reliance, whereas experts apply efficient schemas that support scrutiny but can limit openness to new insights. Importantly, the same modality of expression—SHAP's polished plots—triggers scrutiny for experts but causes misplaced trust for novices. This duality shows that interpretability cannot be designed as one-size-fits-all: its effectiveness depends on who is using it and how. We argue that data science practices can benefit from this duality as a means of social redundancy for appropriate reliance. Finally, we present ideas for design scaffolds that nudge novices toward selective, critical use, while giving experts tools that integrate smoothly into established workflows.

8 Generative AI Disclosure

ChatGPT was used for table formatting, title/subtitle suggestions, grammar and style editing, summary suggestions for the abstract and conclusion, and condensing some sections. No content from generative AI was included verbatim. The authors reviewed and extensively edited these suggestions.

9 Acknowledgements

We would like to thank our reviewers for their constructive comments. We are grateful to everyone in the GroupLens research lab, especially Lana Yarosh, for their feedback and support. We also want to thank the data scientists who participated in our study.

References

[1] Ashraf Abdul, Christian Von Der Weth, Mohan Kankanhalli, and Brian Y Lim. 2020. COGAM: measuring and moderating cognitive load in machine learning model explanations. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–14.

[2] Mark S Ackerman, Juri Dachtera, Volkmar Pipek, and Volker Wulf. 2013. Sharing knowledge and expertise: The CSCW view of knowledge management. *Computer Supported Cooperative Work (CSCW)* 22, 4 (2013), 531–573.

[3] Robin S Adams, Jennifer Turns, and Cynthia J Atman. 2003. What could design learning look like. In *Expertise in Design: Design Thinking Research Symposium*, Vol. 6. Citeseer.

[4] Elena Agapie, Ravi Karkar, Tricia Aung, Eleanor R Burgess, Munyaradzi Joel Chinguwa, Andrea K Graham, Predrag Klasnja, Aaron Lyon, Terika McCall, Sean A Munson, et al. 2024. Conducting Research at the Intersection of HCI and Health: Building and Supporting Teams with Diverse Expertise to Increase Public Health Impact. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–6.

[5] Oscar Alvarado and Annika Waern. 2018. Towards algorithmic experience: Initial efforts for social media contexts. In *Proceedings of the 2018 chi conference on human factors in computing systems*. 1–12.

[6] Karl Anders Ericsson. 2008. Deliberate practice and acquisition of expert performance: a general overview. *Academic emergency medicine* 15, 11 (2008), 988–994.

[7] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2022. Machine bias. In *Ethics of data and analytics*. Auerbach Publications, 254–264.

[8] Theo Araujo, Natali Helberger, Sanne Kruikemeier, and Claes H De Vreese. 2020. In AI we trust? Perceptions about automated decision-making by artificial intelligence. *AI & society* 35, 3 (2020), 611–623.

[9] Hilary Arksey and Lisa O'malley. 2005. Scoping studies: towards a methodological framework. *International journal of social research methodology* 8, 1 (2005), 19–32.

[10] Cynthia J Atman, Justin R Chimka, Karen M Bursic, and Heather L Nachtmann. 1999. A comparison of freshman and senior engineering design processes. *Design studies* 20, 2 (1999), 131–152.

[11] Alan Baddeley. 1976. The psychology of memory. (1976).

[12] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S Lasecki, Daniel S Weld, and Eric Horvitz. 2019. Beyond accuracy: The role of mental models in human-AI team performance. In *Proceedings of the AAAI conference on human computation and crowdsourcing*, Vol. 7. 2–11.

[13] 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*. 1–16.

[14] Alejandro Barredo Arrieta, Siham Tabik, Salvador García López, Daniel Molina Cabrera, Francisco Herrera Triguero, Natalia Ana Díaz Rodríguez, et al. 2019. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. (2019).

[15] Divya Vohra Behl and Susan Ferreira. 2014. Systems thinking: An analysis of key factors and relationships. *Procedia Computer Science* 36 (2014), 104–109.

[16] Astrid Bertrand, Rafik Belloum, James R Eagan, and Winston Maxwell. 2022. How cognitive biases affect XAI-assisted decision-making: A systematic review. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. 78–91.

[17] Astrid Bertrand, James R Eagan, and Winston Maxwell. 2023. Questioning the ability of feature-based explanations to empower non-experts in robo-advised financial decision-making. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 943–958.

[18] Jessica Y Bo, Sophia Wan, and Ashton Anderson. 2025. To rely or not to rely? evaluating interventions for appropriate reliance on large language models. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–23.

[19] Sarah Brayne. 2021. *Predict and surveil: Data, discretion, and the future of policing*. Oxford University Press.

[20] 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. *Proceedings of the ACM on Human-computer Interaction* 5, CSCW1 (2021), 1–21.

[21] Zana Buçinca, Siddharth Swaroop, Amanda E Paluch, Finale Doshi-Velez, and Krzysztof Z Gajos. 2025. Contrastive explanations that anticipate human misconceptions can improve human decision-making skills. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–25.

[22] Ángel Alexander Cabrera, Erica Fu, Donald Bertucci, Kenneth Holstein, Ameet Talwalkar, Jason I Hong, and Adam Perer. 2023. Zeno: An interactive framework for behavioral evaluation of machine learning. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–14.

[23] 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.

[24] 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.

[25] Roberta Calderwood, Gary A Klein, and Beth W Crandall. 1988. Time pressure, skill, and move quality in chess. *The American journal of psychology* (1988), 481–493.

[26] Paul D Callister. 2009. Thinking like a research expert: Schemata for teaching complex problem-solving skills. *Legal Reference Services Quarterly* 28, 1-2 (2009), 31–51.

[27] Astrid Carolus, Martin J Koch, Samantha Straka, Marc Erich Latoschik, and Carolin Wienrich. 2023. MAILS-Meta AI literacy scale: Development and testing of an AI literacy questionnaire based on well-founded competency models and psychological change-and meta-competencies. *Computers in Human Behavior: Artificial Humans* 1, 2 (2023), 100014.

[28] 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.

[29] Diogo V Carvalho, Eduardo M Pereira, and Jaime S Cardoso. 2019. Machine learning interpretability: A survey on methods and metrics. *Electronics* 8, 8 (2019), 832.

[30] Noah Castelo, Maarten W Bos, and Donald R Lehmann. 2019. Task-dependent algorithm aversion. *Journal of marketing research* 56, 5 (2019), 809–825.

[31] Kathy Charmaz. 2008. Reconstructing grounded theory. *The SAGE handbook of social research methods* (2008), 461–478.

[32] Kathy Charmaz, Liska Belgrave, et al. 2012. Qualitative interviewing and grounded theory analysis. *The SAGE handbook of interview research: The complexity of the craft* 2 (2012), 347–365.

[33] Neil Charness. 1976. Memory for chess positions: Resistance to interference. *Journal of Experimental Psychology: Human Learning and Memory* 2, 6 (1976), 641.

[34] Neil Charness. 1981. Search in chess: Age and skill differences. *Journal of Experimental Psychology: Human Perception and Performance* 7, 2 (1981), 467.

[35] William G Chase and K Anders Ericsson. 1982. Skill and working memory. In *Psychology of learning and motivation*. Vol. 16. Elsevier, 1–58.

[36] William G Chase, Don R Lyon, and K Anders Ericsson. 1981. Individual differences in memory span. In *Intelligence and learning*. Springer, 157–162.

[37] William G Chase and Herbert A Simon. 1973. The mind's eye in chess. In *Visual information processing*. Elsevier, 215–281.

[38] William G Chase and Herbert A Simon. 1973. Perception in chess. *Cognitive psychology* 4, 1 (1973), 55–81.

[39] Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 785–794.

[40] 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*. 1–12.

[41] Michelene TH Chi. 2013. Knowledge structures and memory development. In *Children's thinking*. Psychology press, 73–96.

[42] Michelene TH Chi, Paul J Feltovich, and Robert Glaser. 1981. Categorization and representation of physics problems by experts and novices. *Cognitive science* 5, 2 (1981), 121–152.

[43] Michelene T.H. Chi and Robert Glaser. 1982. Final report: knowledge and skill difference in novice and experts. *Contract* 00014 (1982).

[44] Chun-Wei Chiang, Zhuoran Lu, Zhuoyan Li, and Ming Yin. 2023. Are two heads better than one in ai-assisted decision making? comparing the behavior and performance of groups and individuals in human-ai collaborative recidivism risk assessment. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–18.

[45] Kiroong Choe, Seokhyeon Park, Seokweon Jung, Hyeok Kim, Ji Won Yang, Hwajung Hong, and Jinwook Seo. 2024. Supporting novice researchers to write literature review using language models. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–9.

[46] Angèle Christin. 2018. Counting clicks: Quantification and variation in web journalism in the United States and France. *Amer. J. Sociology* 123, 5 (2018), 1382–1415.

[47] Wesley M Cohen, Daniel A Levinthal, et al. 1990. Absorptive capacity: A new perspective on learning and innovation. *Administrative science quarterly* 35, 1 (1990), 128–152.

[48] Juliet M Corbin and Anselm Strauss. 1990. Grounded theory research: Procedures, canons, and evaluative criteria. *Qualitative sociology* 13, 1 (1990), 3–21.

[49] Nigel Cross, Henri Christiaans, and Kees Dorst. 1994. Design expertise amongst student designers. *Journal of Art & Design Education* 13, 1 (1994), 39–56.

[50] Robert L Cross and Sam Israelit. 2009. *Strategic learning in a knowledge economy*. Routledge.

[51] Francisco Cruz and Tania Lombrozo. 2025. How laypeople evaluate scientific explanations containing jargon. *Nature Human Behaviour* (2025), 1–16.

[52] Gloria Dall'Alba and Jörgen Sandberg. 2006. Unveiling professional development: A critical review of stage models. *Review of educational research* 76, 3 (2006), 383–412.

[53] Valdemar Danry, Pat Pataranutaporn, Yaoli Mao, and Pattie Maes. 2023. Don't just tell me, ask me: Ai systems that intelligently frame explanations as questions improve human logical discernment accuracy over causal ai explanations. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–13.

[54] Jeffrey Dastin. 2022. Amazon scraps secret AI recruiting tool that showed bias against women. In *Ethics of data and analytics*. Auerbach Publications, 296–299.

[55] Richard Lee Davis, Thiemo Wambsganss, Wei Jiang, Kevin Gonyop Kim, Tanja Käser, and Pierre Dillenbourg. 2024. Fashioning creative expertise with generative AI: Graphical interfaces for design space exploration better support ideation than text prompts. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–26.

[56] Ankolika De and Zhicong Lu. 2024. # PoetsOfInstagram: Navigating The Practices And Challenges Of Novice Poets On Instagram. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–16.

[57] Adriaan D De Groot. 1946. *Thought and choice in chess*. Amsterdam University Press.

[58] Anurag Deep, Rumana Pathan, and Ritayan Mitra. 2018. Comparing Experts' Systems thinking skill across contexts. In *2018 IEEE Tenth International Conference on Technology for Education (T4E)*. IEEE, 154–157.

[59] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. 2015. Algorithm aversion: people erroneously avoid algorithms after seeing them err. *Journal of experimental psychology: General* 144, 1 (2015), 114.

[60] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).

[61] Hubert Dreyfus and Stuart E Dreyfus. 1986. *Mind over machine*. Simon and Schuster.

[62] Rudresh Dwivedi, Devam Dave, Het Naik, Smiti Singhal, Rana Omer, Pankesh Patel, Bin Qian, Zhenyu Wen, Tejal Shah, Graham Morgan, et al. 2023. Explainable AI (XAI): Core ideas, techniques, and solutions. *Comput. Surveys* 55, 9 (2023), 1–33.

[63] Dennis E Egan and Barry J Schwartz. 1979. Chunking in recall of symbolic drawings. *Memory & cognition* 7, 2 (1979), 149–158.

[64] 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 2024 CHI Conference on Human Factors in Computing Systems*. 1–32.

[65] Noel Entwistle. 2022. Research into learning and teaching in universities: A view from the past towards an uncertain future. In *Student support services*. Springer, 13–33.

[66] Adrian Erasmus, Tyler DP Brunet, and Eyal Fisher. 2021. What is interpretability? *Philosophy & Technology* 34, 4 (2021), 833–862.

[67] Michael Eraut. 2007. Learning from other people in the workplace. *Oxford review of education* 33, 4 (2007), 403–422.

[68] Michael Eraut. 2010. Knowledge, working practices, and learning. *Learning through practice: Models, traditions, orientations and approaches* (2010), 37–58.

[69] Karl Anders Ericsson and Jacqui Smith. 1991. *Toward a general theory of expertise: Prospects and limits*. Cambridge University Press.

[70] Theodore Evans, Carl Orge Retzlaff, Christian Geißler, Michaela Kargl, Markus Plass, Heimo Müller, Tim-Rasmus Kiehl, Norman Zerbe, and Andreas Holzinger. 2022. The explainability paradox: Challenges for xAI in digital pathology. *Future Generation Computer Systems* 133 (2022), 281–296.

[71] Mingming Fan, Xianyou Yang, TszTung Yu, Q Vera Liao, and Jian Zhao. 2022. Human-ai collaboration for UX evaluation: effects of explanation and synchronization. *Proceedings of the ACM on human-computer interaction* 6, CSCW1 (2022), 1–32.

[72] Shi Feng and Jordan Boyd-Graber. 2019. What can ai do for me? evaluating machine learning interpretations in cooperative play. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 229–239.

[73] Paul M Fitts and Michael I Posner. 1967. Human performance. (1967).

[74] James Fleck. 1998. Expertise: knowledge, power and tradeability. In *Exploring expertise: Issues and perspectives*. Springer, 143–171.

[75] Peter Frensch and Robrt J Sternberg. 1987. Expertise and knowledge modification-when bridge isnt bridge anymore. (1987).

[76] Daniel J Garland and John R Barry. 1991. Cognitive advantage in sport: The nature of perceptual structures. *The American Journal of Psychology* (1991), 211–228.

[77] Isabel Gauthier and Michael J Tarr. 2002. Unraveling mechanisms for expert object recognition: bridging brain activity and behavior. *Journal of Experimental Psychology: Human Perception and Performance* 28, 2 (2002), 431.

[78] 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.

[79] 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*. 1–13.

[80] Dennis H Holding and H Douglas Pfau. 1985. Thinking ahead in chess. *The American journal of psychology* (1985), 271–282.

[81] Lu Hong and Scott E Page. 2004. Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences* 101, 46 (2004), 16385–16389.

[82] Shenda Hong, Daoxin Yin, Gongzheng Tang, Tianfan Fu, Liantao Ma, Junyi Gao, Mengling Feng, Mai Wang, Yu Yang, Fei Wang, et al. 2024. Artificial Intelligence and Data Science for Healthcare: Bridging Data-Centric AI and People-Centric Healthcare. In *Proceedings*

of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 6720–6721.

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

[84] Marie Hornberger, Arne Bewersdorff, and Claudia Nerdel. 2023. What do university students know about Artificial Intelligence? Development and validation of an AI literacy test. *Computers and Education: Artificial Intelligence* 5 (2023), 100165.

[85] Chung-Ching Huang and Erik Stolterman. 2014. Temporal anchors in user experience research. In *Proceedings of the 2014 conference on Designing interactive systems*. 271–274.

[86] Edwin Hutchins. 1995. *Cognition in the Wild*. MIT press.

[87] Hyunseung Hwang, Andrew Bell, Joao Fonseca, Venetia Pliatsika, Julia Stoyanovich, and Steven Euijong Whang. 2025. Shap-based explanations are sensitive to feature representation. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*. 1588–1601.

[88] Yoo Hyun Jeong, Sunghyun Hwang, and Dong-Kyu Chae. 2024. HiLite: Hierarchical Level-implemented Architecture Attaining Part-Whole Interpretability. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. 983–993.

[89] Nathanael Jo, Sina Aghaei, Jack Benson, Andres Gomez, and Phebe Vayanos. 2023. Learning optimal fair decision trees: Trade-offs between interpretability, fairness, and accuracy. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. 181–192.

[90] Daniel Kahneman. 2002. Maps of bounded rationality: A perspective on intuitive judgement and choice. (2002).

[91] Harmanpreet Kaur, Eytan Adar, Eric Gilbert, and Cliff Lampe. 2022. Sensible AI: Re-imagining interpretability and explainability using sensemaking theory. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 702–714.

[92] Harmanpreet Kaur, Matthew R Conrad, Davis Rule, Cliff Lampe, and Eric Gilbert. 2024. Interpretability Gone Bad: The Role of Bounded Rationality in How Practitioners Understand Machine Learning. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW1 (2024), 1–34.

[93] 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*. 1–14.

[94] Os Keyes. 2018. The misgendering machines: Trans/HCI implications of automatic gender recognition. *Proceedings of the ACM on human-computer interaction* 2, CSCW (2018), 1–22.

[95] Been Kim, Cynthia Rudin, and Julie Shah. 2014. The bayesian case model: A generative approach for case-based reasoning and prototype classification. *Advances in neural information processing systems* 27 (2014).

[96] 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*. 2390–2395.

[97] 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*. 1–14.

[98] Maya Krishnan. 2020. Against interpretability: a critical examination of the interpretability problem in machine learning. *Philosophy & Technology* 33, 3 (2020), 487–502.

[99] 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*. IEEE, 3–10.

[100] Marianne LaFrance. 1989. The quality of expertise: implications of expert-novice differences for knowledge acquisition. *ACM SIGART Bulletin* 108 (1989), 6–14.

[101] Isaac Lage, Emily Chen, Jeffrey He, Menaka Narayanan, Been Kim, Samuel J Gershman, and Finale Doshi-Velez. 2019. Human evaluation of models built for interpretability. In *Proceedings of the AAAI conference on human computation and crowdsourcing*, Vol. 7. 59–67.

[102] 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*. 29–38.

[103] Himabindu Lakkaraju, Stephen H Bach, and Jure Leskovec. 2016. Interpretable decision sets: A joint framework for description and prediction. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1675–1684.

[104] Jill H Larkin, John McDermott, Dorothea P Simon, and Herbert A Simon. 1980. Models of competence in solving physics problems. *Cognitive science* 4, 4 (1980), 317–345.

[105] Richard P Larrick and Daniel C Feiler. 2015. Expertise in decision making. *The Wiley Blackwell handbook of judgment and decision making* 2 (2015), 696–721.

[106] Bryan R Lawson. 1979. Cognitive strategies in architectural design. *Ergonomics* 22, 1 (1979), 59–68.

[107] Benjamin Letham, Cynthia Rudin, Tyler H McCormick, and David Madigan. 2015. Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. (2015).

[108] Qingzi 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*. 1–15.

[109] 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*. 2119–2128.
[110] Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. 2020. Explainable ai: A review of machine learning interpretability methods. *Entropy* 23, 1 (2020), 18.
[111] Xiao-Yang Liu, Hongyang Yang, Jiechao Gao, and Christina Dan Wang. 2021. FinRL: Deep reinforcement learning framework to automate trading in quantitative finance. In *Proceedings of the second ACM international conference on AI in finance*. 1–9.
[112] Gordon D Logan. 1988. Toward an instance theory of automatization. *Psychological review* 95, 4 (1988), 492.
[113] Jennifer M Logg, Julia A Minson, and Don A Moore. 2019. Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes* 151 (2019), 90–103.
[114] Tania Lombrozo. 2006. The structure and function of explanations. *Trends in cognitive sciences* 10, 10 (2006), 464–470.
[115] Antoine Louis, Gijs van Dijck, and Gerasimos Spanakis. 2024. Interpretable long-form legal question answering with retrieval-augmented large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 22266–22275.
[116] Scott M. Lundberg and contributors. 2025. *SHAP: SHapley Additive exPlanations Documentation.* https://shap.readthedocs.io/en/latest/ Version 0.47.1.
[117] Scott M Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. 2020. From local explanations to global understanding with explainable AI for trees. *Nature machine intelligence* 2, 1 (2020), 56–67.
[118] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems* 30 (2017).
[119] Yuzhen Luo, Kurt Becker, John Gero, Idalis Villanueva Alarcon, and OENARDI Lawanto. 2021. Systems thinking in engineering design: Differences in expert vs. novice. *Int. J. Eng. Educ* 37, 5 (2021), 1398–1413.
[120] Yi Luo, Huan-Hsin Tseng, Sunan Cui, Lise Wei, Randall K Ten Haken, and Issam El Naqa. 2019. Balancing accuracy and interpretability of machine learning approaches for radiation treatment outcomes modeling. *BJR| Open* 1, 1 (2019), 20190021.
[121] Daphne Maurer, Richard Le Grand, and Catherine J Mondloch. 2002. The many faces of configural processing. *Trends in cognitive sciences* 6, 6 (2002), 255–260.
[122] 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 (2019), 1–23.
[123] Katherine B McKeithen, Judith S Reitman, Henry H Rueter, and Stephen C Hirtle. 1981. Knowledge organization and skill differences in computer programmers. *Cognitive Psychology* 13, 3 (1981), 307–325.
[124] Bahar Memarian and Tenzin Doleck. 2024. Data science pedagogical tools and practices: A systematic literature review. *Education and information technologies* 29, 7 (2024), 8179–8201.
[125] George A Miller. 1956. The magical number seven, plus. *Minus Two* (1956).
[126] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence* 267 (2019), 1–38.
[127] Tim Miller. 2023. Explainable ai is dead, long live explainable ai! hypothesis-driven decision support using evaluative ai. In *Proceedings of the 2023 ACM conference on fairness, accountability, and transparency*. 333–342.
[128] Sarah Milne. 2024. *AI tools show biases in ranking job applicants' names according to perceived race and gender.* University of Washington. https://www.washington.edu/news/2024/10/31/ai-bias-resume-screening-race-gender/ Accessed: 2025-04-29.
[129] Naomi Miyake and Donald A Norman. 1978. To Ask a Question, One Must Know Enough to Know What Is Not Known. Report No. 7802. (1978).
[130] Raha Moraffah, Mansooreh Karami, Ruocheng Guo, Adrienne Raglin, and Huan Liu. 2020. Causal interpretability for machine learning-problems, methods and evaluation. *ACM SIGKDD Explorations Newsletter* 22, 1 (2020), 18–33.
[131] Elia Morgulev, Ofer H Azar, and Ronnie Lidor. 2018. Sports analytics and the big-data era. *International Journal of Data Science and Analytics* 5, 4 (2018), 213–222.
[132] William James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. 2019. Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences* 116, 44 (2019), 22071–22080.
[133] Yuri Nakao, Simone Stumpf, Subeida Ahmed, Aisha Naseer, and Lorenzo Strappelli. 2022. Toward involving end-users in interactive human-in-the-loop AI fairness. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 12, 3 (2022), 1–30.
[134] Mitchell J Nathan, Walter Kintsch, and Emilie Young. 1992. A theory of algebra-word-problem comprehension and its implications for the design of learning environments. *Cognition and instruction* 9, 4 (1992), 329–389.
[135] Allen Newell, Herbert Alexander Simon, et al. 1972. *Human problem solving*. Vol. 104. Prentice-hall Englewood Cliffs, NJ.
[136] Janni Nielsen, Torkil Clemmensen, and Carsten Yssing. 2002. Getting access to what goes on in people's heads? Reflections on the think-aloud technique. In *Proceedings of the second Nordic conference on Human-computer interaction*. 101–110.
[137] Jakob Nielsen and Rolf Molich. 1990. Heuristic evaluation of user interfaces. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 249–256.

[138] Taehyung Noh, Haein Yeo, Myungjin Kim, and Kyungsik Han. 2023. A study on user perception and experience differences in recommendation results by domain expertise: the case of fashion domains. In *Extended abstracts of the 2023 CHI conference on human factors in computing systems*. 1–7.
[139] Harsha Nori, Samuel Jenkins, Paul Koch, and Rich Caruana. 2019. InterpretML: A unified framework for machine learning interpretability. *arXiv preprint arXiv:1909.09223* (2019).
[140] Mahsan Nourani, Chiradeep Roy, Jeremy E Block, Donald R Honeycutt, Tahrima Rahman, Eric Ragan, and Vibhav Gogate. 2021. Anchoring bias affects mental model formation and user reliance in explainable AI systems. In *Proceedings of the 26th International Conference on Intelligent User Interfaces*. 340–350.
[141] Chinasa T Okolo. 2023. Navigating the Limits of AI Explainability: Designing for Novice Technology Users in Low-Resource Settings. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. 959–961.
[142] Yitao Peng, Lianghua He, Die Hu, Yihang Liu, Longzhen Yang, and Shaohua Shang. 2024. Decoupling Deep Learning for Enhanced Image Recognition Interpretability. *ACM Transactions on Multimedia Computing, Communications and Applications* 20, 10 (2024), 1–24.
[143] 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*. 1–52.
[144] Charvi Rastogi, Yunfeng Zhang, Dennis Wei, Kush R Varshney, Amit Dhurandhar, and Richard Tomsett. 2022. Deciding fast and slow: The role of cognitive biases in ai-assisted decision-making. *Proceedings of the ACM on Human-computer Interaction* 6, CSCW1 (2022), 1–22.
[145] Eyal M Reingold, Neil Charness, Richard S Schultetus, and Dave M Stampe. 2001. Perceptual automaticity in expert chess players: Parallel encoding of chess relations. *Psychonomic Bulletin & Review* 8, 3 (2001), 504–510.
[146] Judith S Reitman. 1976. Skilled perception in Go: Deducing memory structures from inter-response times. *Cognitive psychology* 8, 3 (1976), 336–356.
[147] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1135–1144.
[148] Don Rogers and John A Sloboda. 2013. *The acquisition of symbolic skills*. Vol. 22. Springer Science & Business Media.
[149] Andrew Ross and Finale Doshi-Velez. 2018. Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.
[150] William B. Rouse. 2003. Engineering complex systems: Implications for research in systems engineering. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 33, 2 (2003), 154–156.
[151] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence* 1, 5 (2019), 206–215.
[152] Cynthia Rudin, Chaofan Chen, Zhi Chen, Haiyang Huang, Lesia Semenova, and Chudi Zhong. 2022. Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistic Surveys* 16 (2022), 1–85.
[153] Hicham Sadok, Fadi Sakka, and Mohammed El Hadi El Maknouzi. 2022. Artificial intelligence and bank credit analysis: A review. *Cogent Economics & Finance* 10, 1 (2022), 2023262.
[154] 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*. 240–251.
[155] Morgan Klaus Scheuerman, Jacob M Paul, and Jed R Brubaker. 2019. How computers see gender: An evaluation of gender classification in commercial facial analysis services. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–33.
[156] Vera Schmitt, Luis-Felipe Villa-Arenas, NIls Feldhus, Joachim Meyer, Robert P Spang, and Sebastian Möller. 2024. The role of explainability in collaborative human-AI disinformation detection. In *Proceedings of the 2024 ACM conference on fairness, accountability, and transparency*. 2157–2174.
[157] Donald A Schön. 1988. Designing: Rules, types and worlds. *Design studies* 9, 3 (1988), 181–190.
[158] Edward Segel and Jeffrey Heer. 2010. Narrative visualization: Telling stories with data. *IEEE transactions on visualization and computer graphics* 16, 6 (2010), 1139–1148.
[159] Phoebe Sengers, Kirsten Boehner, Shay David, and Joseph'Jofish' Kaye. 2005. Reflective design. In *Proceedings of the 4th decennial conference on Critical computing: between sense and sensibility*. 49–58.
[160] Ben Shneiderman. 2003. The eyes have it: A task by data type taxonomy for information visualizations. In *The craft of information visualization*. Elsevier, 364–371.
[161] Herbert A Simon. 1990. Bounded rationality. In *Utility and probability*. Springer, 15–18.
[162] Herbert A Simon and Kevin Gilmartin. 1973. A simulation of memory for chess positions. *Cognitive psychology* 5, 1 (1973), 29–46.
[163] Siddharth Singh, Mayank Kaushik, Ambuj Gupta, and Anil Kumar Malviya. 2019. Weather forecasting using machine learning techniques. In *Proceedings of 2nd international conference on advanced computing and software engineering (ICACSE)*.
[164] Venkatesh Sivaraman, Leigh A Bukowski, Joel Levin, Jeremy M Kahn, and Adam Perer. 2023. Ignore, trust, or negotiate: understanding clinician acceptance of AI-based treatment recommendations in health care. In *Proceedings of the 2023 CHI Conference on Human Factors*

in Computing Systems. 1–18.
[165] Alison Smith-Renner, Ron Fan, Melissa Birchfield, Tongshuang Wu, Jordan Boyd-Graber, Daniel S Weld, and Leah Findlater. 2020. No explainability without accountability: An empirical study of explanations and feedback in interactive ml. In *Proceedings of the 2020 chi conference on human factors in computing systems*. 1–13.
[166] Robert J Sternberg. 2014. *Advances in the Psychology of Human Intelligence: Volume 4*. Psychology Press.
[167] Gregor Stiglic, Primoz Kocbek, Nino Fijacko, Marinka Zitnik, Katrien Verbert, and Leona Cilar. 2020. Interpretability of machine learning-based prediction models in healthcare. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 10, 5 (2020), e1379.
[168] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*. PMLR, 3319–3328.
[169] Harini Suresh, Steven R Gomez, Kevin K Nam, and Arvind Satyanarayan. 2021. Beyond expertise and roles: A framework to characterize the stakeholders of interpretable machine learning and their needs. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–16.
[170] Annalisa Szymanski, Brianna L Wimer, Oghenemaro Anuyah, Heather A Eicher-Miller, and Ronald A Metoyer. 2024. Integrating expertise in llms: crafting a customized nutrition assistant with refined template instructions. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–22.
[171] James W Tanaka and Marjorie Taylor. 1991. Object categories and expertise: Is the basic level in the eye of the beholder? *Cognitive psychology* 23, 3 (1991), 457–482.
[172] Shelley E Taylor and John D Winkler. 1980. The Development of Schemas. (1980).
[173] Suzanne Tolmeijer, Markus Christen, Serhiy Kandul, Markus Kneer, and Abraham Bernstein. 2022. Capable but amoral? Comparing AI and human expert collaboration in ethical decision making. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–17.
[174] John Toner, Barbara Gail Montero, and Aidan Moran. 2015. The perils of automaticity. *Review of General Psychology* 19, 4 (2015), 431–442.
[175] 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 (2023), 1–38.
[176] Alfredo Vellido. 2020. The importance of interpretability and visualization in machine learning for applications in medicine and health care. *Neural computing and applications* 32, 24 (2020), 18069–18083.
[177] Luis Von Ahn and Laura Dabbish. 2008. Designing games with a purpose. *Commun. ACM* 51, 8 (2008), 58–67.
[178] James F Voss, Terry R Greene, Timothy A Post, and Barbara C Penner. 1983. Problem-solving skill in the social sciences. In *Psychology of learning and motivation*. Vol. 17. Elsevier, 165–213.
[179] Jindong Wang, Haoliang Li, Haohan Wang, Sinno Jialin Pan, and Xing Xie. 2023. Trustworthy machine learning: Robustness, generalization, and interpretability. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 5827–5828.
[180] Xinru Wang and Ming Yin. 2021. Are explanations helpful? a comparative study of the effects of explanations in ai-assisted decision-making. In *Proceedings of the 26th International Conference on Intelligent User Interfaces*. 318–328.
[181] Elizabeth Anne Watkins. 2023. Face Work: A Human-Centered Investigation into Facial Verification in Gig Work. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–24.
[182] John Broadus Watson. 1930. Behaviorism, rev. (1930).
[183] Karl E. Weick. 1995. *Sensemaking in organizations*. Sage Publications.
[184] Mark Weiser and Joan Shertz. 1983. Programming problem representation in novice and expert programmers. *International Journal of Man-Machine Studies* 19, 4 (1983), 391–398.
[185] James Wexler, Mahima Pushkarna, Tolga Bolukbasi, Martin Wattenberg, Fernanda Viégas, and Jimbo Wilson. 2019. The what-if tool: Interactive probing of machine learning models. *IEEE transactions on visualization and computer graphics* 26, 1 (2019), 56–65.
[186] A Mark Williams and Paul R Ford. 2008. Expertise and expert performance in sport. *International Review of Sport and Exercise Psychology* 1, 1 (2008), 4–18.
[187] Chengshuo Xia, Tian Min, Daxing Zhang, and Congsi Wang. 2024. Understanding the Needs of Novice Developers in Creating Self-Powered IoT. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–17.
[188] Seyma Yucer, Furkan Tektas, Noura Al Moubayed, and Toby Breckon. 2024. Racial bias within face recognition: A survey. *Comput. Surveys* 57, 4 (2024), 1–39.
[189] Yunfeng Zhang, Q Vera Liao, and Rachel KE 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.

A Intake Survey for Proficiency-based Recruitment

For basic eligibility requirements, participants were asked to confirm that they were: (1) at least 18 years old, and (2) current residents of the United States. This was followed by questions about their educational background and professional experience, such as current field of study or employment and highest degree obtained. Then the participants were asked about their experience with ML and interpretability tools (1–7 Likert scale), including the extent of ML integration in their job roles (1–7 Likert scale). Hornberger et al. [84]'s validated objective AI Literacy Scale asked 31 multiple choice questions, each falling into one of these categories: Recognizing AI, Understanding Intelligence, Ethics, Decision-Making, Data Literacy, and Understanding Model Behavior.

B Interview and Think-Aloud Protocol

B.1 Introduction

The research team and project were introduced to participants as follows:

> *We are a team of researchers at the University of Minnesota. We are conducting this study to observe how expertise influences people's understanding of AI outputs and explanations. We will share an ML task with you to understand how you approach it. We will be here to address any questions or concerns. We will ask you to turn on your video and also share your screen once we share the study documents with you.*

Consent. Participants were asked to turn on video for privacy and data security protocols and were asked: *"Do you consent to this video being recorded?"*

Introducing the Task. The following instructions were read to participants:

> *Before we start talking, we are sharing a Google Colab notebook with you. This notebook already has a dataset, model, and interpretability tool setup. Feel free to run the notebook. We will also ask you to share your screen with us as you work with the notebook. We would like for you to conduct an exploratory data analysis on this dataset: walk us through how you would evaluate the data and model, use the interpretability tool to support that exploration, and feel free to write your own code as well. We are also sharing a tutorial with you that describes the dataset, the model and the SHAP plots that we use in the Colab notebook. We have some multiple-choice questions based on the data that we would ultimately like you to answer, but let us spend these first few minutes on just the exploration and not focus on the MCQs just yet.*

B.2 Pre-Study Interview

The following was read to participants: *"We will ask you a couple of questions to establish your background in machine learning. You don't have to go into too much detail here; we're simply looking for high-level context on your ML experience."*

(1) Can you describe your background in machine learning, including any formal training or practical experience you've had?
(2) How many years of industry experience do you have?
(3) How long have you been doing machine learning?
(4) What is your current job role or field of study and how does it relate to ML?
(5) *[Verify current degree.]*
(6) Imagine you are working on a new dataset. Can you walk us through your process of making decisions about data selection, pre-processing, and model choice?
(7) What types of checks or validations would you perform?

(8) How would you evaluate whether a model is ready for deployment?

B.3 During the Task

The following instructions were read to participants:

> *We would like you to go over the notebook and the tutorial and explore its contents. While you do this exploration, we want you to think aloud and share with us what you are thinking as you look at our model.*

Since this is a think-aloud protocol, the researcher nudged participants to verbalize their thinking, reasoning, and decision-making processes as they worked through the data analysis task.

B.4 Post-Study Interview

(1) What was your strategy when working with the SHAP plots?
(2) Did SHAP visualizations directly influence your MCQ choices?
(3) Can you describe a specific decision or MCQ answer where the SHAP visualizations directly influenced your choice?
(4) Did any SHAP plot feel unnecessary?
(5) Do you consider yourself a novice or an expert? How would you evaluate your expertise on this task in particular? How do you think you did?

C Data Science Task and Artifacts

C.1 Dataset

We used a publicly available tabular dataset commonly employed for ML classification tasks: the Titanic Survival dataset.[4] This dataset contains information about passengers aboard the Titanic, which sank in 1912 after colliding with an iceberg. Each observation corresponds to a unique passenger and includes input features for passenger class, sex, age, number of siblings or spouses aboard, number of parents or children aboard, ticket number, fare, cabin number, and port of embarkation. The outcome label—survival—is binary, indicating whether or not a passenger survived. This dataset is widely recognized, well-structured, and contains a mix of categorical and numerical features and a binary classification label, making it ideal for our data science task with a set of diversely proficient participants.

C.2 SHAP Tutorial

This tutorial was created based on official SHAP documentation,[5] with a focus on plots used for tabular datasets instead of the wider applications of the official documentation. The tutorial included an overview of the SHAP framework, the dataset and black-box model (XGBoost) used for explanation, and examples of all SHAP visualizations included in the study, along with their interpretations. To avoid exposing participants to plots they would later be tested on, these visualizations were not updated to be based on the Titanic dataset; we re-used the official SHAP plots based on the UCI Income dataset.

C.3 Model and Interpretability Tool

We trained an XGBoost classifier on the Titanic dataset to predict passenger survival based on available features. XGBoost is a gradient boosting algorithm known for its efficiency, scalability, and high predictive performance [39]. While XGBoost is highly effective, it operates as a black-box model, making it difficult to understand how it

[4] https://www.kaggle.com/competitions/titanic/
[5] https://shap.readthedocs.io/en/latest/

arrives at predictions. To interpret the model's predictions, we used SHAP, which is a post-hoc explanation tool.[6] SHAP uses a game-theoretic approach that assigns each feature a contribution score to explain its impact on individual predictions [118]. SHAP provides both local and global interpretability. Global explanations help analyze overall feature importance and trends across all data points. Local explanations describe individual predictions by showing how much each feature contributes to the predicted outcome of a specific instance.

C.4 Colab Notebook

The Google Colab Notebook was designed for exploratory data analysis, model validation, and interpretability. The notebook was organized into sections typical to data science tasks: imports, dataset exploration, model training, various visualizations for SHAP-based explanations, and MCQs.

The imports sections contained the imports for necessary libraries for data handling, visualization, and explainability. The dataset section consisted an overview of the Titanic dataset, including the input features and target variable. We loaded the dataset and performed basic pre-processing on the data such as handling categorical variables and dropping irrelevant columns. Following data preparation, an XGBoost model was trained on the Titanic dataset, and its performance was assessed through accuracy on both training and test sets and log loss on the test data.

In the next section, a SHAP explainer was instantiated, and various visualizations were produced for global and local interpretability. First, a SHAP global summary bar plot was used to highlight the overall importance of features across the dataset. Second, SHAP dependence scatter plots were included to show interactions between two features and their effect on model output. For local explanations, two plots were provided: the SHAP local bar plot and the local decision plot. Additionally, a counterfactual table was introduced, comparing a selected data point with similar instances to show how minimal changes in the feature values would lead to a different model prediction.

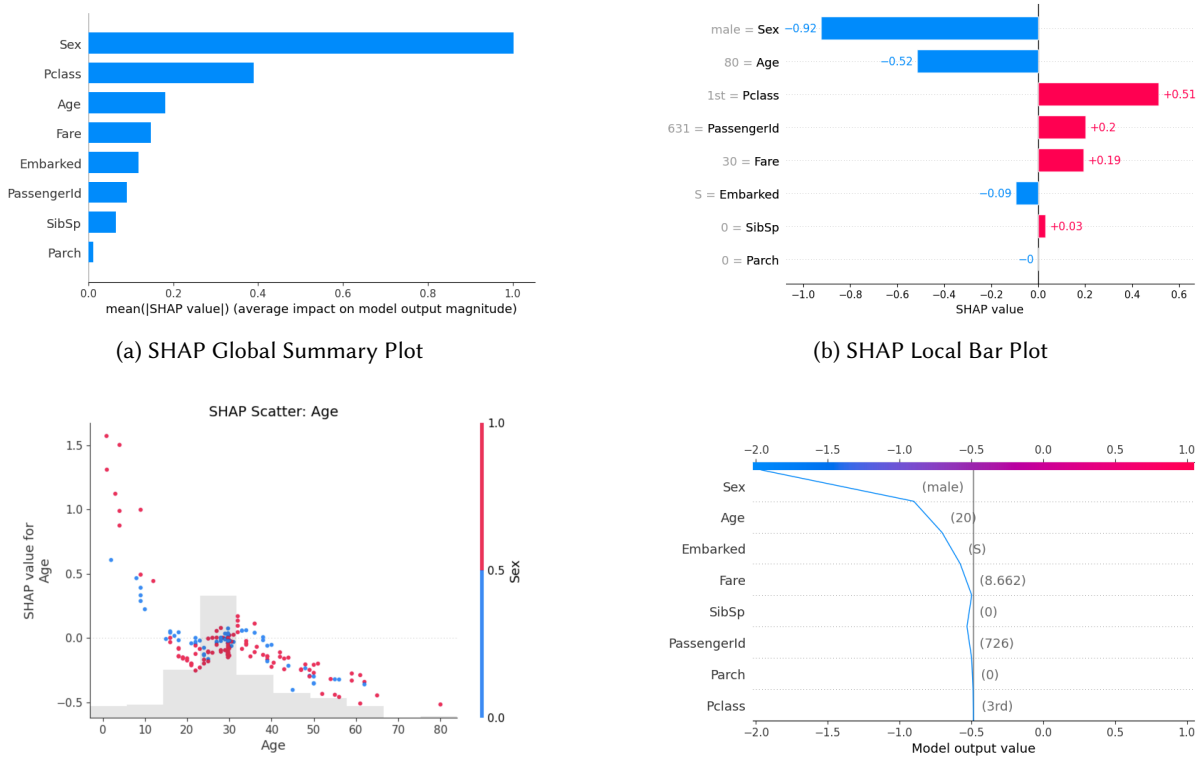
C.5 Post-Exploration Multiple Choice Questions (MCQs)

Post-exploration, participants answered 10 MCQs while thinking aloud. The MCQ section was structured to evaluate participants' comprehension and usage of the dataset, model, and SHAP outputs.

Each question fell into one or more of the following categories:

- Global Explanation: Questions that asked participants to identify globally important features from SHAP summary plots.
- Local Explanation: Items focused on individual predictions, prompting users to interpret local SHAP values or decision plots.
- Plot Recognition and Interpretation: These tested the ability to distinguish between SHAP visualization types (e.g., bar vs. dependence plots, global vs. local explanations) and interpret colors, value ranges, and patterns.
- What-if Reasoning: Participants analyzed how small changes in feature values would impact model predictions.
- Misclassification Analysis: Some questions presented misclassified instances and required reasoning about possible causes using SHAP insights.
- Comparative reasoning: Participants were asked to compare two or more data points to identify differences in predicted outcomes.
- Understanding Model Behavior: Questions that tested participant's ability to reason about the overall decision-making logic of the model, such as identifying feature interactions or recognizing when the model might be overfitting or biased.

[6] https://shap.readthedocs.io/en/latest/index.html



(a) SHAP Global Summary Plot

(b) SHAP Local Bar Plot

(c) SHAP Dependence Scatter Plot

(d) SHAP Local Decision Plot

Fig. 2. SHAP visualizations generated for the Titanic survival dataset using an XGBoost model. SHAP is a post-hoc explainability tool that explains the predictions of black-box models.

The answer options for MCQs were designed to capture different reasoning strategies, from correct interpretations to plausible but incorrect ones. Several questions in the assessment tested the ability of participants to synthesize information from multiple visual cues and integrate that with their own reasoning. These questions evaluated whether participants relied mainly on the visualizations or used their intuition and prior knowledge to form conclusions. In some cases, participants had to identify patterns across various plots or make predictions based on subtle interactions between features. This tested how well they could connect the dots, recognize underlying patterns, and how they developed reasoning from both visual data and their mental models. Participants also rated their confidence in each answer on a 7-point Likert scale, where 1 represented 'very uncertain' and 7 indicated 'very confident'. This scale helped assess the level of certainty in their interpretations.

pID	Gender	Age	Education	Occupation	Job Experience (months)	Machine Learning Knowledge (1–7)	Interpretability Tool Familiarity (1–7)	Objective Literacy Score (%)
Experts								
P1	Male	24	MS	Teaching Assistant	14	6	5	93.55
P2	Male	26	MS	ML/AI Researcher	14	5	2	87.10
P3	Male	30	PhD	ML/AI Researcher	5	6	4	80.65
P9	Male	23	PhD	ML/AI Researcher	15	5	3	90.32
P11	Male	23	MS	Data Analyst	18	5	5	77.42
P14	Male	29	MS	Data Analyst	48	5	1	45.16
P15	Male	27	PhD	Data Scientist	42	6	7	87.10
P18	Male	30	PhD	Electrical Engineer	65	6	7	77.00
P19	Male	27	MS	Data Scientist	5	5	3	77.00
P20	Male	27	MS	Data Scientist	20	5	5	93.00
P21	Female	23	MS	Student	6	6	4	77.00
P23	Male	23	MS	ML/AI Researcher	24	5	6	90.32
Novices								
P4	Male	21	BS	Software Dev.	4	6	4	54.84
P5	Female	23	MS	Student	3	5	5	51.61
P6	Female	25	MS	Engineer	5	3	1	74.19
P7	Female	26	MS	Research Eng.	38	4	4	77.42
P8	Male	29	MS	Student	13	3	1	38.71
P10	Male	23	PhD	ML/AI Researcher	3	5	2	77.42
P12	Male	25	BS	Student	4	5	3	83.87
P13	Female	23	MS	Teaching Assistant	15	4	1	74.19
P16	Female	22	MS	Student	5	5	2	64.52
P17	Female	25	MS	Software Dev.	29	4	1	77.42
P22	Male	28	MS	Engineer	40	6	6	93.55

Table 1. Overview of participant demographics captured via the intake survey.

		Novices	Experts
Cognitive Processes	**Chunking**	Store small chunks of isolated information that is more easily accessible. However, it is harder to determine the relevance of when to access it.	Store larger chunks of related information about a task that can be referenced collectively in the future.
	Pattern Matching	Do not identify shared cues within or across information types. This makes pattern matching and quick solutions cognitively harder to identify.	Recognize meaningful patterns that let them anticipate connections between information components and apply appropriate known solutions.
	System Thinking	Focus on elements in isolation, which can make it hard to see how variables interact as a whole.	Connect elements into a coherent whole, understanding how variables in a system influence one another.
	Memory and Recall	Use short-term memory when doing tasks because they do not have sufficient indexed information to recall.	Retrieve meaningful patterns relevant to the task from long-term memory.
Knowledge Development and Organization	**Schemas and Indexing**	Rely on less connected schemas, keeping concepts isolated. Approach problems more openly without using preconceptions.	Use more complex schemas to represent interconnected concepts. Work harder to filter out irrelevant details at runtime.
	Domain Knowledge and Categorization	Categorize problems by domain-specific features and rely on memorized solutions.	Categorize problems by domain-agnostic principles or algorithms, enabling them to generalize across tasks and domains.
	Types of Knowledge	Use textbook based knowledge given they are less experienced. This gives them a strong foundation but less flexibility in unfamiliar contexts.	Combine theory, skills and personal experience , using them flexibly depending on the context.
Runtime Task Strategies	**Automaticity**	Complete a task step-by-step without making memory- or intuition-based leaps.	Complete a task automatically by relying on historical experiences with similar setup. However, they struggle when routines or patterns are disrupted.
	Problem Framing and Scoping	Struggle to identify solvable framings of the problem via scoping. Trying to tackle everything at once can hinder their progress.	Scope and set up the problem before attempting to solve it, and continue to iteratively scope as needed.
	Reasoning Style	Primarily deductive, relying on rules and formulae before acting, which constrains flexibility.	Primarily inductive, generating and refining solutions, supported by stronger self-evaluation
	Search Strategies	Rely on broad trial and error searches which might make their process less efficient, but can also help discover important alternatives and edge cases.	Conduct structured, knowledge-driven searches that help them access relevant information quickly for efficient problem-solving.

Table 2. Summary of the key information from our Framework of Expertise, grounded in a scoping review of prior work on the topic from non-technical fields.

Open Code	Axial Code	Theme
Experts: *wants to know the hypothesis being tested in the study; would first identify study objective before data explorations; would use lightgbm instead; wondering why we chose to include specific pdps; what is the study hoping to achieve*	**Experts:** Evaluated the study through the lens of their own workflows and assumptions; scoped the purpose and formed hypotheses before engaging with the data	Experts worked top-down, beginning with objectives, forming hypotheses, and evaluating specifics against that initial understanding. Novices worked bottom-up, relying on model predictions and SHAP to guide their understanding rather than forming their own hypotheses first.
Novices: *starts by data exploration; will do data cleaning based on data type; building intuition based on shap plots; gradually developing understanding of plots; referred back to exact plot details*	**Novices:** Followed the task structure as given, following step-by-step; adopted an exploratory approach; built their understanding of the data incrementally through SHAP outputs rather than independent analysis	
Novices: *questioned colors and what they mean; tried to understand values and ranges of axes; interpreting legends; interpreting marks and channels*	**Novices:** Focused on decoding the visual elements of SHAP plots: colors, axes, marks; memorized specific plot details and recalled them for answering post-study questions (memory/retrieval strategy)	Experts and novices allocated attention at different levels. Experts skimmed plot-level details and instead focused on meta-level questions about SHAP's mechanics. Novices decoded each plot thoroughly: its colors, axes, and marks, memorizing these details for later recall.
Experts: *curious about hyperparameters; how to merge information from different plots; shap plots useful for quick visual takeaway; would come back to plots if needed but will skim for now*	**Experts:** Allocated attention strategically, examining one representative plot and skimming similar ones (attention strategy); redirected attention to meta-level concerns: how SHAP values are calculated, how to synthesize across plots, what the tool's assumptions are	
Experts: *not enough data to make a prediction using shap; wrote their own code; more confidence in traditional methods from years of experience; trust SHAP when it maps to domain knowledge; SHAP adds a layer of less transparency; SHAP overfits on anomalies and biases; want to see SHAP background calculations; not use SHAP for making conclusions*	**Experts:** Trust in SHAP was conditional on alignment with their domain knowledge and prior experience; perceived SHAP as adding opacity rather than removing it—would have liked to see background SHAP calculations; when SHAP conflicted with their understanding, experts trusted standard data science methods	Experts treated SHAP as a tool to be checked, conditioning trust on alignment with domain knowledge, and defaulting to their own methods when outputs contradicted their understanding. Novices treated SHAP as a source of ground truth, accepting its outputs as accurate and sufficient for drawing conclusions.
Novices: *SHAP plots give the exact accurate information needed; relied more on plots than intuition; trust shap plots to answer MCQs; appreciates accuracy of shap plots*	**Novices:** Treated SHAP outputs as accurate and sufficient, not questioning the assumptions behind the visualizations; based their post-study assessments solely on SHAP plots, using them as the primary evidence for reasoning	

Table 3. Open codes, axial codes, and emerging themes from our grounded theory analysis.