Evaluating Interpretation Risks in Interactive What-If Scenarios and Mitigating with HypotheX

Yulia Grushetskaya[1][0009−0003−5385−2928], Mike Sips[1][0000−0003−3941−7092], and Thomas Kosch[2][0000−0001−6300−9035]

[1] GFZ Helmholtz Centre for Geosciences
[2] Humboldt University Berlin, Department of Computer Science, Germany

Abstract. Human-centered XAI conceptualizes explanations as external representations of model behavior intended to support users' reasoning and sensemaking, rather than as self-contained artifacts that directly convey understanding. Because their interpretation is shaped by users' cognitive processes and contextual circumstances, understanding emerges through active engagement in formulating, testing, and refining hypotheses. In exploratory settings, where relevant questions and factors are not fully specified in advance, such engagement relies on iterative questioning, experimentation, and revision, typically mediated through interactive visualization.

Hypothetical Scenario (HS)–based XAI methods operationalize this exploratory mode of explanation by enabling users to manipulate input conditions and observe resulting changes in model outputs. Through what-if testing, HS externalizes users' reasoning processes by instantiating hypotheses as concrete interactions whose consequences become observable.

Exploratory interaction, however, is not cognitively neutral. Prior work in cognitive science and human–computer interaction shows that reasoning during exploration is systematically shaped by cognitive biases such as anchoring, availability, and salience. In HS-based exploration, these biases can influence scenario selection, input sampling, outcome comparison, and the sequencing of interactions, leading to incomplete or misleading interpretations even when explanations are technically faithful.

Such biases do not manifest as incorrect explanations, but as characteristic interaction behaviors that signal elevated risks of misinterpretation. Accordingly, this paper examines bias-related interpretation risks in HS-based XAI and evaluates whether targeted interface design safeguards can reduce bias-prone interaction patterns. We introduce HypotheX, an interactive HS-based XAI system, and show that interface-level interventions can mitigate some—though not all—such risks.

Keywords: Human-Centred XAI · Cognitive Biases · Hypothetical Scenario (HS) based methods · Interactive XAI Tools

1 Introduction

Modern machine learning (ML) models are complex systems whose decisions are often difficult to interpret. Explaining such decisions requires balancing two competing objectives: providing explanations that faithfully reflect model behavior and ensuring they are understandable and usable for human stakeholders [1, 2]. To address this tension, XAI research has increasingly adopted human-centered approaches that account for how users perceive, reason about, and interact with explanations [3–6].

A central implication of human-centered XAI is that understanding is not guaranteed by the mere availability of an explanation artifact. Explanations function as external representations that support users' reasoning and sensemaking, and their interpretation depends on cognitive processes and analytical context. In exploratory settings, where relevant questions and factors are not specified in advance, understanding emerges through active engagement, such as forming hypotheses, testing them through interaction, and revising them in response to the model's outputs.

Exploratory interaction, however, is not cognitively neutral. Prior work in XAI shows that explanations can be interpreted through the lens of cognitive heuristics and biases, leading to systematic misjudgments even when explanations are technically faithful [12, 13]. In interactive settings, where users actively probe model behavior, these interpretive tendencies can shape how scenarios are selected, tested, and integrated into a final explanation.

In this paper, we examine such interpretive risks in interactive, scenario-based XAI methods, focusing on what-if exploration, which we refer to more formally as Hypothetical Scenario (HS) approaches. While HS-based methods enhance accessibility and user-driven inquiry, they also enable biases to shape interpretation. Without guidance, users may focus on salient instances, overweight recent outcomes, or form confident explanations after limited exploration.

We address two research questions: (1) how bias-related interpretation risks manifest as observable interaction patterns during HS-based exploration, and (2) to what extent interface design can reduce their prevalence. We derive bias-compatible interaction patterns from prior literature and quantify them using interaction-based metrics. To evaluate design-level mitigation, we introduce HypotheX (Sect. 4.2), an HS-based XAI system that integrates interface safeguards to prevent misleading interpretations. A comparative experiment with the What-If Tool shows that HypotheX users exhibit broader and more balanced exploration patterns, indicating reduced bias-compatible behaviors.

2 Related Work

2.1 Human-Centered Explanation

Human-centered XAI (HCXAI) has evolved into several distinct directions, each addressing different aspects of how users perceive and evaluate explanations.

A foundational line of work for HCXAI, grounded in social science theory, focused on the kinds of explanations users perceive as meaningful. Miller [3] identified core explanatory properties such as contrastivity (explaining why one outcome occurred instead of another), selectivity (highlighting the most relevant factors), and causality (revealing underlying mechanisms). These insights inspired a range of XAI methods based on local examples and counterfactuals [7, 8, 22], enabling users to explore and compare specific instances as a basic mechanism for understanding model behavior.

Another important line of research shifted attention to users' intent, background, and context, acknowledging that the effectiveness of an explanation depends on how well it aligns with users' goals, expertise, and decision-making roles [9, 4, 10]. This perspective led to the development of explanation strategies and interface designs tailored to diverse user groups—such as model developers, lay users, or decision makers [5, 11]. Interactivity emerged as a key mechanism for personalization, allowing users to probe, modify, and refine explanations according to their specific needs and understanding [4, 5].

Another perspective frames explanation as a cognitive process, shaped by users' interpretive effort, heuristics, and reasoning strategies. In this view, users actively construct meaning through exploration and selective attention, rather than passively receiving information. This approach draws attention to cognitive biases—such as anchoring, availability, and overconfidence—that can skew interpretation even when explanations are technically correct [12, 13, 19]. As a result, there is increasing focus on interface-level interventions designed to support more deliberate and reflective reasoning during explanation use.

Recent works also underscore the importance of exposing the limitations and uncertainties of AI systems [14]. Rather than concealing failures, this approach—sometimes called seamful XAI—advocates for explanations that reveal where models are uncertain.

HypotheX builds on key explanatory principles from prior research. Drawing on contrastive and selective approaches, we use a hypothetical scenario (HS) format to let users explore localized input–output relationships. From user-centered perspectives, we prioritize interactivity and agency, allowing individuals to engage with explanations at their own pace and toward their own goals. In line with seamful design, our interface exposes prediction probabilities, enabling users to observe both classification outcomes and shifts in model confidence; revealing behavior beyond predictions.

However, our primary focus aligns with the view of explanation as an active, cognitively mediated process. Understanding model behavior through interactive exploration—especially when users define and try to achieve their own explanation goals—is a mentally demanding task. Like other complex cognitive activities, this process is highly vulnerable to heuristics and biases, which can distort interpretation even when explanations are technically sound. We specifically investigate how biases such as anchoring, availability, and overconfidence manifest in scenario-based exploration, and we design safeguards to help users recognize and mitigate associated risks.

2.2 Cognitive Bias in HS-Based XAI Methods

The influence of cognitive biases on human decision-making has been studied extensively in psychology [15, 16], human–computer interaction, and, more recently, explainable AI. Early XAI research often focused on how explanations increase trust and usability, but growing empirical evidence has shown that they can also lead users to overtrust [13], misinterpret, or oversimplify model behavior [17]. These challenges have prompted a shift in human-centered XAI: from designing "intuitive" explanations to critically analyzing how people actually engage with them under cognitive constraints [12]. Bove et al. [18] frame these risks as a form of "user-specific failure", distinct from flaws in model logic or technical presentation. Even when explanations are technically correct, users may draw inaccurate or incomplete conclusions due to cognitive shortcuts, prior beliefs, or exploratory habits.

A growing body of XAI research has documented numerous cognitive biases that can arise in interactive settings. Bertrand et al. [12] and Bove et al. [18] provide typologies showing how such biases can distort explanation use and model interpretation. Biases frequently identified in XAI studies include: **anchoring bias** (overweighting early or visually salient information) [12, 18]; **availability bias** (relying on recently seen or easily accessed information) [16, 12]; **illusion of explanatory depth** (overestimating understanding after minimal exposure) [19, 17]; **confirmation bias** (favoring evidence that supports prior beliefs) [4, 18]; and **overreliance bias** (placing undue trust in model outputs, even when incorrect) [13, 20].

Cognitive biases are latent mental processes whose behavioral expression is highly context-dependent, influenced by explanation format, task framing, and user expertise. As noted by Bove et al. [18], attributing observed behavior directly to cognitive bias risks overinterpretation or misattribution. At the same time, prior work suggests that when the interaction context is sufficiently bounded, user behavior can still provide a measurable record of bias-compatible reasoning. In visual analytics, Wall et al. [24] formalized this idea by proposing coverage- and distribution-based metrics to detect behavioral indicators of bias from interaction logs, while explicitly arguing against a one-to-one mapping between biases and metrics — a stance we adopt as well. Our work extends this foundation in two directions specific to XAI: we specialize the framework to hypothesis-driven HS-based exploration, where the cognitive unit is not data point selection but hypothesis formation and testing, and we move beyond detection to evaluate whether targeted interface design can reduce bias-compatible behavior.

Building on this perspective, we focus on bias manifestations: concrete, recurring interaction patterns that are theoretically linked to cognitive biases and observable during HS-based exploration. These patterns provide a principled way to relate cognitive theory to empirical interaction data and to analyze where explanation use is susceptible to misinterpretation. Accordingly, we adopt a cautious attribution stance, treating manifestations as indicators of increased interpretation risk rather than as direct evidence of internal cognitive states.

3 Methodology

To study interpretive risks arising from cognitive biases in scenario-based XAI, we propose a systematic framework centered on observable interaction behavior. Rather than inferring users' internal cognitive states or judging the correctness of their understanding, it focuses on interaction patterns compatible with biased reasoning and elevated risks of misinterpretation. The framework connects cognitive bias theory to observable behavior and to interface design choices that shape exploration. Its aim is not to diagnose bias or guarantee correct interpretation, but to detect, characterize, and reduce behavioral risk factors in interactive explanation use.

Operationalizing this perspective requires a systematic procedure that links abstract cognitive biases to observable properties of exploratory behavior. To this end, we adopt a stepwise framework that (i) scopes relevant cognitive biases based on theoretical and contextual considerations, (ii) translates these biases into empirically observable interaction patterns, (iii) defines quantitative metrics to detect such patterns, and (iv) evaluates whether targeted interface design choices can reduce their occurrence. Importantly, each step is designed to remain agnostic to users' internal cognitive states and to operate solely on interaction traces and interface mechanisms. The framework thus provides a systematic means to analyze and mitigate bias-compatible interaction behavior in interactive XAI without assuming access to cognition or ground truth understanding. Our application of the framework is shown in Fig. 1.

3.1 Bias Selection

We focus on cognitive biases and interpretive errors that were identified empirically as relevant in interactive, HS-based XAI.

Anchoring bias refers to the tendency to rely disproportionately on initial information when making judgments or decisions [15]. In HS-based tools, exploration often begins from default instances, central projections, or visually salient regions of the data space. Such starting points can anchor subsequent exploration, leading users to sample only narrow regions of the input space, terminate exploration prematurely, or generalize from unrepresentative cases.

Availability bias describes the tendency to rely more strongly on information that is easier to recall, more recent, or more perceptually salient [16]. In HS-based exploration, interaction unfolds over time and across multiple scenarios. Memory limitations and visual salience can therefore bias interpretation toward recent tests or prominent instances, distorting how cumulative evidence is integrated in the final explanation.

Overreliance denotes excessive trust in model outputs or explanatory visualizations, even when their scope, assumptions, or limitations are not fully understood [20]. In HS-based XAI, this may manifest as interpreting visible patterns in low-dimensional projections as faithful representations of the model's internal reasoning, without systematically probing alternative scenarios or considering unobserved dependencies.

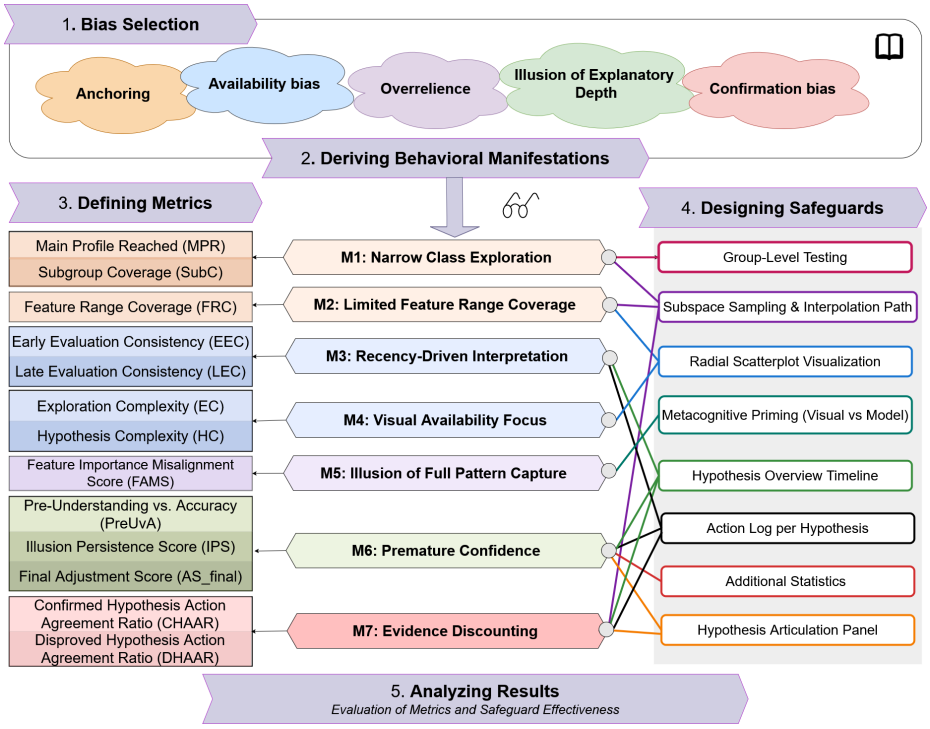

Fig. 1. Bias-aware evaluation framework. The framework links cognitive bias selection (Step 1) to behavioral manifestations (Step 2), operational metrics (Step 3), safeguard design (Step 4), and analysis of metrics and safeguard effectiveness (Step 5).

Illusion of Explanatory Depth (IOED) is a metacognitive error in which individuals overestimate their understanding of complex systems after limited interaction [19, 17]. Interactive XAI interfaces, by providing immediate and seemingly transparent feedback, may foster premature confidence and early termination of exploration before important edge cases or failure modes are encountered.

Confirmation bias refers to the tendency to seek, interpret, and evaluate evidence in ways that favor existing beliefs or hypotheses [21]. In HS-based workflows, this can manifest as selectively testing scenarios that support an emerging interpretation while discounting or downplaying contradictory outcomes, leading to asymmetric hypothesis evaluation.

These biases were selected because they are well described in the scientific literature, likely to influence exploratory interaction in HS-based XAI, and suitable for operationalization through interaction traces. Other biases—such as representativeness or framing effects—are also relevant to explanation use, but fall outside the scope of this study due to limitations in detectability or

experimental control. Their investigation remains an important direction for future work.

3.2 Deriving Behavioral Manifestations

Behavioral manifestations do not constitute direct evidence of cognitive bias. The same interaction pattern may arise from rational, task-adaptive strategies, and any given cognitive bias may produce multiple behavioral expressions depending on context. Accordingly, we treat manifestations not as diagnostic indicators of internal cognitive states, but as interaction-level signals of elevated interpretation risk. This cautious attribution stance allows us to relate cognitive theory to empirical interaction data without inferring mental processes. From this perspective, we derive seven recurring manifestations (M1–M7) that capture observable patterns of interaction associated with elevated interpretation risks during HS-based explanation.

M1: Narrow Class Exploration (Anchoring-Compatible Pattern). The user's exploration remains concentrated in one region or visually dominant subgroup of a class, while other regions receive little attention. A plausible explanation is anchoring: once a user starts from a particular cluster or set of instances, subsequent exploration may stay close to that starting point instead of testing whether the same explanation holds across the class. Explanations formed in this way can miss within-class heterogeneity and may not generalize beyond the explored region.

M2: Limited Feature Range Coverage (Anchoring-Compatible Pattern). The user varies feature values only within a narrow interval, usually near default, central, or otherwise salient values in the interface. Here anchoring may appear at the level of feature manipulation, with the current setting acting as a reference point that keeps later adjustments local. This reduces the chance of encountering thresholds, nonlinear responses, or feature interactions outside the sampled range, so the resulting interpretation may capture only locally valid behavior.

M3: Recency-Driven Interpretation (Availability-Compatible Pattern). The user's final interpretation is shaped mainly by recently tested scenarios, whereas earlier interactions play a smaller role in later reasoning. This fits the logic of availability bias: information encountered most recently is often easiest to recall when a judgment is made. The risk is that earlier contradictions or alternative behaviors are not fully integrated, so the final explanation reflects only the latest part of the exploration history.

M4: Visual Availability Focus (Availability-Compatible Pattern). Exploration is driven largely by what is immediately visible or easy to access in the interface, such as default projections, preselected feature axes, or centrally positioned instances. In this case, visually foregrounded information may guide what users attend to and what they consider worth testing. The result can be a narrow exploration pattern in which less salient regions of the data space remain underexamined, and explanations reflect interface visibility more than the broader structure of model behavior.

M5: Illusion of Full Pattern Capture (Overreliance-Compatible Pattern). The user treats visually apparent structures in low-dimensional views, such as clusters, gradients, or boundaries, as if they were sufficient evidence of the model's decision logic, without probing alternative explanations through targeted scenario testing. This is where overreliance becomes relevant: confidence in the displayed output extends beyond what the projection alone can justify. As a result, users may read high-dimensional structure into partial visual evidence and miss dependencies or interactions that are not visible in the current view.

M6: Overestimation of Understanding (Illusion of Explanatory Depth). The user reports strong confidence in an explanation after only limited or unbalanced exploration. In HS-based interaction, a few successful tests may already create a sense of understanding, even when the evidential basis is still thin. Exploration may therefore stop too early, with the user assuming that the model's reasoning has already been sufficiently understood.

M7: Evidence Discounting (Confirmation-Compatible Pattern). During hypothesis testing, the user does not treat supporting and contradicting evidence equally. Confirming outcomes may be accepted quickly, whereas disconfirming results are discounted, postponed, or treated as insufficient. This kind of asymmetry is consistent with confirmation bias, where evidence that supports an emerging belief is often given more weight than evidence that challenges it. In practice, this can keep exploration aligned with an initial hypothesis and reduce attention to plausible alternatives.

3.3 Metric Development

To analyze the behavioral manifestations introduced above, we define a set of interaction-based and self-report–based metrics that operationalize observable aspects of these patterns.

Two principles guide metric development. First, metrics are manifestation-specific rather than bias-specific: each metric targets a concrete behavioral property (e.g., range coverage, temporal weighting of outcomes, or alignment between perceived and model-derived importance). Second, we do not assume that all proposed metrics are equally sensitive or robust. Especially for manifestations related to hypothesis formation, confidence calibration, or evidence integration, it remains open whether short-term interaction data are sufficient to produce stable signals. Detectability and robustness are therefore part of the evaluation.

We distinguish between coverage-oriented metrics (e.g., M1, M2), which describe how broadly users traverse the input or class space, and reasoning-oriented metrics (e.g., M3, M6, M7), which describe how outcomes are weighted and incorporated into users' final explanations over time. The former are expected to be more directly shaped by interface properties, whereas the latter may reflect stronger individual differences.

The following metrics operationalize each manifestation and define the expected signal indicating elevated interpretive risk. These expected signals serve as hypotheses for later analysis rather than as validation claims.

To illustrate the metrics concretely, we refer throughout to examples from one experimental test case, the Duck class of the Animal Hobbies dataset (features include Swimming, Gardening, Running, Art, and Detective Stories).

M1: Narrow Class Exploration (Anchoring Bias). M1 captures whether participants explored representative and diverse regions of a class.

We operationalize this using k-means clustering within each class to identify internal subgroups. The resulting clusters represent different feature profiles within the same class and thus capture its internal variation.

Main Profile Reached (MPR) indicates whether a participant tested at least one instance belonging to the largest within-class cluster. This cluster represents the most common feature profile of the class.

In the Duck class, many instances share a pattern of low Gardening and high Swimming. If a participant repeatedly tests Duck cases but never probes this dominant feature profile, they may form an explanation without having accessed the class core. In such cases, MPR remains 0.

Subgroup Coverage (SubC) measures how many distinct within-class clusters the participant tested relative to the total number of clusters identified for that class. Duck contains multiple subtypes—for example, strongly Swimming-driven Ducks and others combining moderate Swimming with Art or Detective Stories. Testing only one such subtype results in low SubC, even if many instances were examined.

Expected signal. Low MPR and low SubC indicate exploration confined to a limited portion of the class structure rather than engagement with its representative and diverse regions.

M2: Limited Feature Range Coverage (Anchoring Bias). M2 concerns how far participants move within a feature's value range during exploration.

We measure this using **Feature Range Coverage (FRC)**. For each feature, we identify the lowest and highest values a participant tested in the interaction log. The distance between these two values is compared to the full range of that feature in the dataset. The resulting proportions are then aggregated across features.

In the Duck class, a participant may keep *Detective Stories* near a mid value (e.g., 5–8) and observe no change in prediction, leading them to conclude that the feature is irrelevant. However, they never test the ends of the scale (e.g., 0–1 or 13–14), where the prediction probability may shift under certain feature combinations. Although several adjustments are made, only a small segment of the 0–14 range is covered.

Expected signal. Low FRC indicates that testing was confined to a limited portion of the available feature range rather than spanning a broader interval.

M3: Recency-Driven Interpretation (Availability Bias). M3 captures whether participants rely more strongly on recent testing outcomes than on earlier evidence when accepting or rejecting their hypothesis.

Early Evaluation Consistency (EEC) reflects how strongly the outcomes of the first 30% of hypothesis-testing actions align with the participant's final decision. **Late Evaluation Consistency (LEC)** reflects the same alignment for the last 20% of actions.

For each segment, we compute an aggregate agreement score by comparing the outcome of each test action with the participant's final hypothesis judgment, indicating whether the evidence in that segment was predominantly supportive or contradictory. The metrics, therefore, indicate how much early versus late evidence is reflected in the final judgment.

In the Duck class, several typical instances with low Gardening and high Swimming are first tested and consistently predicted as Duck. Near the end of exploration, an unusual case with moderate Gardening and very high Detective Stories shifts the prediction to Cat.

If the final explanation is based mainly on this last example, Detective Stories appears decisive, while the earlier stable pattern is disregarded. The interpretation would then be driven by the recent outlier rather than the overall evidence. In this case, LEC exceeds EEC.

Expected signal. Recency-driven interpretation is indicated when LEC is systematically higher than EEC, suggesting stronger weighting of recent evidence over earlier observations.

M4: Visual Availability Focus (Availability Bias). M4 concerns cases where hypotheses are shaped mainly by the information that is available during the initial exploration phase, while other parts of the input space receive little attention.

Exploration Complexity (EC) describes how broadly the input space was inspected *before* a hypothesis was stated. It captures how many distinct feature combinations, classes, and regions were visited during this initial exploration phase.

Hypothesis Complexity (HC) describes how structurally rich the stated hypotheses are. It reflects whether hypotheses reference multiple features, use value ranges or thresholds, or consider more than one outcome class.

In the Duck class, reasoning could begin in the Swimming–Gardening projection. Observing that low Gardening and high Swimming often yield Duck predictions, a hypothesis such as "Ducks occur when Swimming is high and Gardening is low" could be formed. If other projections, such as Swimming–Running, are not examined before the hypothesis is stated, interactions that influence boundary cases may remain untested. For example, high Swimming combined with high Running can shift predictions toward Dog. When the explanation remains tied to a single 2D projection without integrating additional feature pairs, HC remains low relative to the available interaction structure.

Expected signal. M4 is expected to appear as low HC values, indicating hypotheses framed around a small set of currently available features. EC serves as an opportunity measure: if initial exposure differs strongly between conditions, corresponding shifts in HC would be expected only if hypothesis structure is primarily driven by what was available during that exposure.

M5: Illusion of Full Pattern Capture (Overreliance Bias). M5 concerns cases where visually clear structure in 2D projections is taken as a direct reflection of the model's decision logic.

To estimate visual salience, we computed silhouette scores for all 2D feature pairs. For each feature, we averaged the silhouette scores across all pairs in which the feature appeared. This yields a ranking of features according to how consistently they produce visually separable clusters in scatter plots. Model importance was derived with SHAP (SHapley Additive exPlanations) values [25].

We then compared participant feature-importance rankings (elicited via questionnaire) to both the visual salience ranking and the SHAP ranking. Alignment was quantified by combining set overlap and rank agreement. The difference between model alignment and visual alignment indicates whether explanations are guided more by the model's actual feature importance or by visually salient patterns, such as apparent cluster separation, in the low-dimensional view.

In the Duck class, Swimming produces strong visual separation in several 2D projections and therefore ranks highly in visual salience. However, SHAP analysis shows that Gardening plays an equally central role in the model's decision logic. If feature-importance judgments prioritize Swimming because its clusters appear in scatter plots, while underweighting Gardening's contribution, the explanation reflects visual structure rather than model structure.

Expected signal. The manifestation is indicated when participant feature rankings align more strongly with visual salience than with model importance, suggesting that visually apparent patterns are treated as complete representations of the model's behavior.

M6: Premature Confidence with Sparse Testing (Illusion of Explanatory Depth). M6 captures discrepancies between subjective confidence and actual task performance.

Participants rated their understanding of the model before and after completing a prediction task. Ratings were collected on a 1–5 Likert scale and normalized to a 0–1 range. Task performance was measured as classification accuracy.

PreUvA (Pre-Understanding vs. Accuracy) quantifies the discrepancy between normalized pre-task confidence and actual accuracy. Positive values indicate overestimation of understanding prior to receiving feedback.

AS(Adjustment Score) quantifies whether post-task confidence adjustment is consistent with task performance. To compute it, we compare the direction and magnitude of the confidence change from pre- to post-task with the participant's accuracy. The score is higher when confidence is reduced after poor performance and lower when confidence remains high or increases despite incorrect performance. Negative values indicate increased confidence despite low accuracy.

For example, if confidence is rated 4 (normalized 0.75) before the task, accuracy is 0.25, and confidence remains at 4 afterward, there is no downward

adjustment despite weak performance. By contrast, if confidence drops from 4 (0.75) to 2 (0.25) after the same low accuracy, AS becomes positive.

These measures describe whether confidence is calibrated to evidential support generated during exploration.

Expected signal. M6 is indicated by positive PreUvA values and low or negative AS, reflecting inflated initial confidence and limited adjustment following task feedback.

M7: Evidence Discounting (Confirmation Bias). M7 captures whether test outcomes are weighted differently when confirming versus rejecting a hypothesis.

CHAAR (Confirmed Hypothesis Action Agreement Ratio) is the proportion of test actions whose outcomes support a hypothesis that has been confirmed.

DHAAR (Disproved Hypothesis Action Agreement Ratio) is the proportion of test actions whose outcomes contradict a hypothesis that has been rejected.

These ratios quantify how strongly final decisions align with the outcomes generated during exploration.

For example, in the Duck class, a rule such as "High Swimming predicts Duck" may be tested multiple times. If several test outcomes support the rule but some contradict it (e.g., high Swimming combined with high Running predicts Dog), the rule may still be confirmed by the user. In this case, CHAAR would be moderate because confirmation occurs despite mixed test outcomes.

By contrast, a competing rule such as "High Art predicts Duck" may be rejected only after repeated contradictory outcomes. A high DHAAR means that a participant rejected the hypothesis only when a large share of the tested outcomes contradicted it.

Expected signal. M7 is indicated when DHAAR exceeds CHAAR, meaning that rejection requires stronger alignment with test outcomes than confirmation.

3.4 Designing Safeguards

To reduce bias-related interpretation risks in HS-based exploration, HypotheX integrates a set of interface safeguards that target specific behavioral manifestations. These safeguards are designed as opportunity structures: they alter what users can see, test, or revisit during interaction, thereby shaping exploration trajectories and supporting reflection.

Importantly, the presence of a safeguard does not guarantee its use, nor does use guarantee mitigation of the targeted interpretation risk. Safeguards vary in how directly they intervene in the exploration process. We therefore distinguish between: (i) *Exploration-structuring safeguards*, which modify default traversal paths or sampling affordances, and (ii) *Reflection-oriented safeguards*, which rely on voluntary engagement and impose additional cognitive effort. We expect these two classes to differ in effectiveness under time pressure and goal-driven tasks.

Exploration-Structuring Safeguards

Radial Scatterplot Visualization. Extends conventional 2D scatterplots by encoding additional feature dimensions radially around each point. This safeguard increases per-state information density and is intended to reduce overreliance on a single pair of axes (M4) and encourage consideration of higher-dimensional interactions (M2). At the same time, richer visualizations can make apparent clusters more visually compelling, which may encourage users to rely on them as if they fully reflected the model's logic (M5).

Group-Level Testing. Allows users to test entire subgroups or clusters rather than isolated instances. This shifts interaction from point-wise to group-wise exploration and is intended to counter narrow class exploration (M1) by promoting representative sampling.

Subspace Sampling and Interpolation Paths. Enable systematic traversal between instances or across defined feature ranges. These operations are designed to reduce anchoring on local regions (M1) and limited feature range coverage (M2). By facilitating generation of intermediate and counter-hypothetical cases, they may also support disconfirmation-oriented testing (M7), though this effect depends on the user.

Reflection-Oriented Safeguards

Additional Statistics(e.g., prediction confidence, feature influence). Expose uncertainty and local model behavior during exploration. These statistics are intended to discourage premature confidence (M6) by highlighting variability and limits of apparent patterns.

Hypothesis Articulation Panel. Prompts users to explicitly state hypotheses before testing and revisit them afterward. This externalization is intended to support metacognitive monitoring (M6) and reduce asymmetric evidence evaluation (M7). Its effectiveness depends on active user engagement.

Action Log and Hypothesis Timeline. Provide persistent access to prior tests, outcomes, and hypothesis revisions. These safeguards are intended to mitigate recency-driven interpretation (M3) and support balanced evidence integration. Because they primarily support retrospective reasoning, their impact may be limited when tasks reward rapid completion.

Metacognitive Priming for Visual Pattern–Model Discrepancies.

Rather than relying solely on interface mechanisms, we explicitly prime participants during onboarding that visually salient structures in low-dimensional projections may not reflect the model's true decision logic. This intervention targets illusion-of-full-pattern-capture risks (M5) at the level of user expectations rather than interaction mechanics.

We explicitly do not assume that all safeguards will be effective; the evaluation tests which manifestations are structurally modifiable by interface design and which are not.

4 User Study Design and Procedure

Our user study investigated whether bias-related manifestations emerge during HS-based exploration of machine learning models and whether interface-level

safeguards influence their occurrence. The What-If Tool served as a baseline HS-based interface without explicit bias-targeted safeguards, while HypotheX implemented the proposed design interventions. Both methods were browser-based and integrated with pre-trained models.

4.1 Setup and Procedure

Participants explored two Random Forest classifiers trained on synthetic datasets comprising three classes and five input features. The datasets were constructed with non-linear decision boundaries and embedded feature interactions to increase interpretive complexity and create conditions under which bias-compatible interaction patterns could arise.

Each participant completed two sessions: one using the What-If Tool with Dataset 1, followed by one using HypotheX with Dataset 2. Each session consisted of a brief instructional video introducing HS-based exploration, four tasks designed to elicit specific manifestations, and post-task questionnaires including open-ended feedback.

Sessions were audio- and screen-recorded using a think-aloud protocol. The recordings were annotated to derive interaction traces, including hypotheses, scenario manipulations, and decision sequences. All encoded and anonymized results, datasets, and models are available on GitHub [3].

4.2 Tools

What-If Tool Developed by Wexler et al. [22], the What-If Tool (WIT) is an open-source interface designed for interactive model probing via feature manipulation and prediction inspection. We employ WIT as a baseline because it provides a standard environment for HS exploration without the inclusion of bias-targeted safeguards. This allows for a controlled comparison to evaluate how HypotheX improves upon basic HS-based analysis.

HypotheX We developed HypotheX to support users in conducting hypothesis-driven exploration of machine learning models, with a specific focus on reducing cognitive biases during HS-based exploration. This section first explains how HypotheX supports HS-based exploration of model explanations, following principles similar to existing methods like the What-If Tool, and then describes the safeguards we introduce to mitigate cognitive biases.

The HypotheX interface is explicitly designed to support the iterative reasoning cycle central to HS-based exploration. This cycle consists of four key stages: (1) hypothesis formulation, (2) targeted testing through scenario manipulation, (3) evidence evaluation, and (4) hypothesis refinement or revision. To support this process, HypotheX is organized into four coordinated views: Exploration Views, Testing Editor, Hypothesis Management, and Hypothesis History. In addition, lightweight statistical metrics provide real-time feedback on tested data ranges, class coverage, and overlap with previous actions.

[3] https://github.com/grushety/HypotheX-User-Study.git

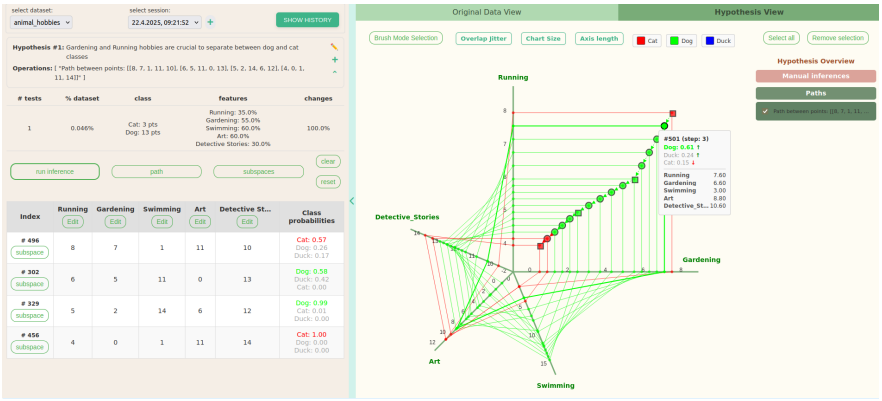

Fig. 2. HypotheX. Top left — Hypothesis configuration panel with summary statistics; bottom left — data editor for modifying feature values; right — hypothesis testing view displaying manipulated instances and their prediction outcomes. The hypothesis testing view presents the input data points using a radial scatterplot, where the user has applied the path interpolation safeguard to explore feature transitions.

Exploration Views. HypotheX separates the exploration process into two phases: pre exploration and evaluation. The Original Data View supports the pre-exploration phase by presenting only the original, unmodified data points and associated model predictions, helping users identify patterns, generate hypotheses, and establish a mental baseline. The Hypothesis View supports the evaluation phase by showing only user-generated or modified data points, allowing users to test assumptions in a controlled, isolated setting. This separation serves two purposes: (1) it preserves the integrity of the original data distribution, reducing the risk of misinterpretation, and (2) it focuses user attention by clearly separating what the user observes from the baseline data versus what the user actively changes or tests. This distinction reduces cognitive overload, helping users see which patterns arise from their manipulations.

Testing Editor. Testing Editor provides tools for users to modify input features, run hypothetical tests, and observe the resulting changes in model predictions, supporting the targeted exploration of 'what if' questions. Users can manually adjust values, apply operations to classes or selected data groups, or use advanced interpolation paths and subspace sampling functions. Note that interpolation paths and subspace sampling are the novel safeguards that we introduce for HS-based exploration. The editor also displays prediction probabilities, enabling users to track how their manipulations affect model certainty and behavior within the Hypothesis View.

Hypothesis Management Panel. Allows users to formulate, record, and organize their working hypotheses, making explicit what they are testing and tracking the assumptions that guide their exploration. The panel helps prevent 'hypothesis

drift', where users unintentionally shift focus without updating their stated assumptions.

History Timeline. Maintains a chronological log of the user's hypotheses tested, explored HSs, and observed results. The timeline is particularly valuable for HS workflows, which often involve backtracking, reconsideration, or transfer of insights between related hypotheses. Users can reload and continue working on any prior hypothesis, enabling iterative refinement over time.

4.3 Participants

We recruited 10 participants with academic backgrounds in science or computer science, most of whom had advanced degrees. A custom questionnaire guided each participant through the entire study.

4.4 HS-Based Tasks and Reasoning Contexts

We designed five HS-based blocks varying exploration freedom, hypothesis specification, and evaluation demands.

Task 1 (Class Profiling) required unconstrained exploration of a target class to identify relevant features and value ranges under time pressure.

Task 2 (Hypothesis Evaluation) involved focused testing of a predefined hypothesis without class discovery.

Task 3 (Free Exploration) required participants to formulate and test their own hypothesis under minimal constraints.

The **Self-Evaluation block** measured confidence adjustment by collecting understanding ratings before and after a prediction task.

Task 4 (Constrained Feature Exploration) required hypothesis construction using features underexplored in Task 1, limiting reliance on familiar dimensions and increasing rejection likelihood.

Together, these tasks define distinct reasoning contexts for observing interaction-based manifestations.

5 Analysis

This section reports results for each manifestation, assessing (i) detectability using the proposed metrics and (ii) the effect of HypotheX safeguards. Because the experiment follows a within-participant design in which each participant interacts with both tools, statistical comparisons evaluate behavioral differences across interface conditions within participants rather than between independent groups. We therefore use repeated-measures ANOVA, which accounts for participant-specific variability. Although the sample size is modest ($n = 10$), such sizes are typical for controlled within-participant studies in visualization and HCI.

5.1 Overall Manifestation Trends

Across manifestations, the results indicate a consistent distinction between exploration-level risks that can be reshaped by interface design and reasoning-level risks that persist despite it.

Interface-modifiable risks: Exploration-level manifestations.

Participants frequently explored narrow regions within classes (M1) and tested features within small value ranges (M2). HypotheX significantly increased both class-level and feature-range coverage, demonstrating that group-level testing, subspace sampling, interpolation paths, and enriched visualization can systematically broaden exploration. The key insight is not that narrow exploration was eliminated—substantial participant-level variation persisted—but that structural safeguards consistently shifted users toward more representative sampling.

M4 and M5 both concern how users rely on information that is immediately available at the point of hypothesis formation or judgment. For M4, Exploration Complexity differed significantly between tools as anticipated: HypotheX's radial visualization reduced sequential navigation needs. Hypothesis Complexity remained low in both conditions, suggesting that participants formulated similarly simple hypotheses regardless of interface differences. One possible explanation is that the short, task-focused setting left limited room for more elaborate hypothesis construction. However, examining EC-HC correspondence within each tool separately reveals no systematic relationship. In both tools, participants with extensive exploration formulated hypotheses of equivalent complexity to those with minimal exploration. The highest HC values did not occur at exploration extremes but at mid-range EC levels. This lack of correspondence between EC and HC within the same tool condition indicates that HC does not reliably capture visual availability bias. If hypotheses were constrained by immediately visible information, broader exploration should have exposed users to more features and yielded more complex hypotheses—yet this pattern did not emerge. Instead, HC appears to reflect individual explanatory style: users who formulated simple hypotheses did so consistently regardless of exploration breadth. Low HC could therefore reflect a focus on visually available information, but also strategic simplification, time constraints, or stable preferences for parsimonious explanations. Without convergent measures to isolate visual availability effects from these alternatives, M4 detection through HC remains limited in this study.

For M5, participants' feature-importance judgments aligned more strongly with visually salient structure than with the model's actual feature importance. As expected, HypotheX amplified this effect through its extended visualization, thereby strengthening the illusion that the visible pattern fully captures the underlying model behavior. Notably, metacognitive priming during onboarding—explicitly warning participants that visual patterns may not reflect model logic—did not prevent this effect. This reveals a fundamental tension: enriching visual information to support exploration can simultaneously increase

overreliance on visually compelling but incomplete projections, even when users are explicitly cautioned against such reliance.

Interface-insensitive risks: Reasoning-level manifestations. Several manifestations showed stable patterns across tools and tasks. Final judgments aligned more strongly with recent evidence than earlier outcomes (M3), and hypothesis rejection required stronger outcome alignment than confirmation (M7: DHAAR > CHAAR). In both cases, the corresponding metrics showed no meaningful shifts under HypotheX safeguards despite the availability of history views and hypothesis timelines. The stability of these asymmetries suggests they reflect general cognitive tendencies in sequential hypothesis testing that are not readily altered by optional retrospective review mechanisms.

M6 (premature confidence) was detectable through participant-level discrepancies between confidence and task accuracy. However, the two tools did not differ clearly in either PreUvA or Adjustment Score. This indicates that while M6 is empirically detectable, it appears largely interface-insensitive under the present design. The significant participant-level variance suggests that confidence calibration reflects stable individual reasoning styles rather than interface-mediated processes.

The findings do not indicate that reflection-oriented safeguards are inherently ineffective. Rather, optional reflection affordances (e.g., history views, hypothesis articulation panels) were systematically underutilized under task pressure. Interaction patterns suggest that participants prioritized forward, hypothesis-testing actions over retrospective evidence integration, leaving recency effects, confirmation asymmetry, and confidence calibration largely unaffected. A plausible explanation is production bias [23], whereby users prioritize actions that directly advance task completion and undervalue reflective activities whose benefits are delayed or indirect. In this setting, reflection-oriented safeguards introduce additional cognitive load without providing immediate task payoff, which reduces their likelihood of use even when they are designed to mitigate reasoning errors. Overall, the results indicate that exploration-structuring safeguards can reshape how users traverse the input space. In contrast, reasoning-level manifestations likely require more directive mechanisms, such as enforced falsification steps, systematic evidence review, or explicit prompts to consider disconfirming evidence, in order to influence belief updating.

More complete metric-level results are provided in the following tables. Aggregated metric values and extended study materials—including annotated interaction traces, metric code, and analysis scripts—are available on GitHub [4].

5.2 Detailed Manifestation Analysis

The following tables summarize the results for each manifestation, reporting metric values, statistical significance, detection validity, and safeguard effects.

[4] https://github.com/grushety/HypotheX-User-Study.git

Table 1. Analysis of anchoring-related manifestations M1 and M2.

	M1: Narrow Class Exploration	**M2: Limited Feature Range Coverage**
Metrics	Main Profile Reached (MPR) Subgroup Coverage (SubC)	Feature Range Coverage (FRC)
Results	- Participant-level values range from very low to near full coverage. - In the What-If Tool, mean MPR (0.39) and mean SubC (0.36) are lower than in HypotheX (0.81 and 0.70). - ANOVA indicates a significant main effect of tool for both metrics ($p < 0.001$).	- Feature Range Coverage is higher under HypotheX (0.71) than under WIT (0.48). - Participant-level values range from 0.21–0.66 (WIT) and 0.33–0.96 (HX). - ANOVA indicates a significant main effect of tool ($p < 0.001$).
Analysis	- Coverage values vary widely across participants, indicating stable differences in exploration breadth rather than task constraints. - The pattern appears across tasks, suggesting a general exploration tendency. - Limited access to dominant and diverse class regions increases the risk that explanations rely on partial class representation.	- Some participants explored only a limited portion of the available feature range. - The wide spread of FRC values indicates stable differences in how broadly feature values are tested. - Restricted range coverage reduces exposure to threshold and non-linear model behavior.
Detection	**Reliable**: Repeated low MPR and SubC values across participants and tasks indicate persistent restricted class-level exploration.	**Reliable**: Repeated low FRC values across participants and tasks indicate persistent narrow value-range exploration.
Safeguards	**Effective**: Higher MPR and SubC under HypotheX, together with a significant main effect of tool, indicate broader class coverage across participants and tasks.	**Effective**: Higher FRC under HypotheX, together with a significant main effect of tool, indicates broader feature value coverage across participants and tasks.

Table 2. Analysis of availability-related manifestations M3 and M4.

	M3: Recency-Driven Interpretation	**M4: Visual Availability Focus**
Metrics	Early Evaluation Consistency (EEC) Late Evaluation Consistency (LEC)	Exploration Complexity (EC) Hypothesis Complexity (HC)
Results	- Across tools and tasks, LEC is consistently higher than EEC (WIT: 0.88 vs. 0.78; HX: 0.84 vs. 0.74). - Both metrics remain high overall, with most participant values between approximately 0.70 and 0.90. - ANOVA shows no significant main effect of tool for EEC ($p = 0.232$) or LEC ($p = 0.233$).	- EC differs between tools (WIT = 5.88; HypotheX = 3.28), with ANOVA indicating a significant effect of tool ($F = 16.999$, $p < 0.001$). - HC remains low across conditions (≈ 0.20–0.23). ANOVA does not indicate a significant effect of tool ($F = 1.517$, $p = 0.222$). - HC varies strongly across participants ($F = 2.087$, $p = 0.040$).
Analysis	- High EEC values indicate that early evidence is incorporated into final judgments. - The consistent LEC > EEC pattern suggests that later evidence exerts stronger influence on the final interpretation. - Because the asymmetry is stable across tools and tasks, it reflects a general weighting tendency rather than an interface-induced effect. - The pattern indicates asymmetric evidence weighting, not rejection of earlier information.	- The tool alters the breadth of interaction states explored prior to hypothesis formulation. - However, differences in exploration breadth are not reflected in differences in hypothesis complexity. - The dissociation between EC and HC suggests that exposure structure and hypothesis articulation are weakly coupled under the present task constraints.
Detection	**Reliable**: Detection is supported because a stable LEC–EEC asymmetry appears across participants and tasks, indicating systematic overweighting of recent evidence in final judgments.	**Limited**: Detection is limited: although exposure breadth varies systematically, hypothesis structure remains stable, providing no convergent evidence of a visual availability effect at the level of hypothesis formulation.
Safeguards	**Ineffective**: HypotheX does not significantly alter EEC, LEC, or their asymmetry, and the recency pattern remains stable across conditions.	**Ineffective**: HypotheX reduces EC while HC remains similar, indicating that extended visualization alters exploration structure without influencing hypothesis formation.

Table 3. Analysis of manifestations M5–M7.

	M5: Illusion of Full Pattern Capture	**M6: Premature Confidence**	**M7: Evidence Discounting**
Metrics	Feature Importance Misalignment Score (FAMS)	Pre-Understanding vs. Accuracy (PreUvA) Adjustment Score (AS)	Confirmed/Disproved Hypothesis Action Agreement Ratio (CHAAR/DHAAR)
Results	- Visual alignment differs between tools ($F(1,18) = 6.18$, $p = 0.0229$). - Model alignment does not show a significant effect of tool ($F(1,18) = 0.75$, $p = 0.3988$). - Model alignment varies significantly across participants ($F(9,10) = 13.98$, $p = 0.0001$).	- Mean accuracy is identical across tools (0.55), no significant tool effect - Mean PreUvA and mean AS are near zero. - Participant-level effects are significant for AS ($p = 0.005$) and show a trend for PreUvA ($p = 0.066$).	- DHAAR exceeds CHAAR across tools and tasks. - No significant tool effect for CHAAR ($p = 0.916$) or DHAAR ($p = 1.000$).
Analysis	- Participants' feature-importance judgments align more closely with visually salient structure than with the model's true feature importance. - Tool differences affect visual alignment (increased by HypotheX). - Strong participant-level variation.	- Confidence–accuracy alignment varies primarily between participants, not tools. - Near-zero tool means indicate no systematic change in confidence calibration at the tool level.	- The asymmetry indicates stronger integration of disconfirming evidence in rejection than confirming evidence in acceptance. - The stability of this pattern across tools, task and participants suggests it reflects a general evaluation tendency.
Detection	**Reliable:** Systematic differences between visual and model alignment indicate consistent reliance on visually salient structure.	**Reliable:** Persistent discrepancies between confidence adjustment and accuracy across participants support the detection of the manifestation.	**Reliable:** The persistent DHAAR > CHAAR asymmetry across participants and tasks indicates systematic differences in evidence integration.
Safeguards	**Adverse effect:** Enhanced visualization in HypotheX increased overreliance on visual features.	**Ineffective:** HypotheX does not alter confidence–accuracy alignment or adjustment patterns.	**Ineffective:** Safeguards didn't reduce CHAAR–DHAAR gap. Optional review and articulation insufficient.

6 Conclusion

We presented a bias-aware evaluation framework for interactive HS-based XAI that studies interpretive risk through observable interaction behavior rather than inferred cognitive states. Using seven manifestations (M1–M7) and corresponding metrics, we showed that several interpretation risks can be detected from interaction traces.

A clear distinction emerged between exploration-level and reasoning-level risks. Exploration-level manifestations (M1, M2) were both measurable and responsive to interface design. Under HypotheX, group-level testing, subspace sampling, and interpolation paths increased class coverage and feature-range coverage, shifting users toward broader and more representative exploration of the input space.

Reasoning-level manifestations (M3, M6, M7) were also detectable, but changed little across tools. Final judgments were shaped more by recent evidence than by earlier outcomes, confidence calibration varied strongly across participants, and rejecting a hypothesis required stronger outcome alignment than confirming one. These patterns remained even when history views and articulation panels were available, suggesting that optional reflection features were not enough to change evidence integration in short, task-focused settings.

M5 highlighted a further tradeoff. The richer visualization in HypotheX broadened feature awareness, but also increased the tendency to treat visually salient structure as if it fully captured the model's logic. In this case, feature-importance judgments followed visual salience more than systematic scenario testing.

Taken together, the results suggest that interface design can change how users explore, but has less influence on how they weigh evidence and update beliefs. Reducing reasoning-level risks will likely require more directive workflow mechanisms, such as enforced falsification steps or structured evidence review, rather than optional reflective features.

This study is limited by its small sample size, short exposure duration, and synthetic task setting. We used synthetic datasets to keep conditions controlled and comparable across tools, but this also reduces ecological realism. The findings should therefore be read as evidence of plausible trends under controlled conditions, not as general proof across domains. Future work should test the framework on real-world data.

By linking cognitive biases to measurable interaction patterns and concrete interface mechanisms, our framework establishes a practical methodology for bias-aware XAI design. This work also demonstrates that supporting accurate interpretation in interactive XAI requires not only expanding what users can explore, but also structuring how exploratory evidence is integrated into explanations.

References

1. Adadi, A., Berrada, M.: Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). IEEE Access **6**, 52138–52160 (2018)
2. Barredo Arrieta, A., et al.: Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges. Inf. Fusion **58**, 82–115 (2020)
3. Miller, T.: Explanation in artificial intelligence: Insights from the social sciences. Artif. Intell. **267**, 1–38 (2019)
4. Wang, D., Yang, Q., Abdul, A., Lim, B.Y.: Designing theory-driven user-centric explainable AI. In: Proceedings of the CHI Conference on Human Factors in Computing Systems, pp. 1–15. ACM (2019)
5. Liao, Q.V., Varshney, K.R.: Human-centered explainable AI (XAI): From algorithms to user experiences. arXiv preprint arXiv:2110.10790 (2021)
6. Ehsan, U., Wintersberger, P., Liao, Q.V., et al.: Human-centered explainable AI (HCXAI): Beyond opening the black-box of AI. In: CHI Conference on Human Factors in Computing Systems Extended Abstracts, Article 109, pp. 1–7. ACM (2022)
7. Wachter, S., Mittelstadt, B., Russell, C.: Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harvard J. Law Technol. **31**, 841 (2018)
8. Mothilal, R.K., Sharma, A., Tan, C.: Explaining machine learning classifiers through diverse counterfactual explanations. In: Proceedings of the Conference on Fairness, Accountability, and Transparency, pp. 607–617. ACM (2020)
9. Liao, Q.V., Gruen, D., Miller, S.: Questioning the AI: Informing design practices for explainable AI user experiences. In: Proceedings of the CHI Conference on Human Factors in Computing Systems, pp. 1–15. ACM (2020)
10. Eiband, M., Schneider, H., Bilandzic, M., Fazekas-Con, J., Haug, M., Hussmann, H.: Bringing transparency design into practice. In: Proceedings of the 23rd International Conference on Intelligent User Interfaces, pp. 211–223. ACM (2018)
11. Ehsan, U., Liao, Q.V., Muller, M., Riedl, M.O., Weisz, J.D.: Expanding explainability: Towards social transparency in AI systems. In: Proceedings of the CHI Conference on Human Factors in Computing Systems, Article 82, pp. 1–19. ACM (2021)
12. Bertrand, A., Belloum, R., Eagan, J.R., Maxwell, W.: How cognitive biases affect XAI-assisted decision-making: A systematic review. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES). ACM (2022)
13. Buçinca, Z., Malaya, M.B., Gajos, K.Z.: To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. Proc. ACM Hum.-Comput. Interact. 5(CSCW1), Article 188 (2021).
14. Ehsan, U., Liao, Q.V., Passi, S., Riedl, M.O., Daumé III, H.: Seamful XAI: Operationalizing seamful design in explainable AI. Proc. ACM Hum.-Comput. Interact. **8**, 1–29 (2024)
15. Tversky, A., Kahneman, D.: Judgment under uncertainty: Heuristics and biases. Science **185**(4157), 1124–1131 (1974)
16. Kahneman, D.: Thinking, Fast and Slow. Farrar, Straus and Giroux, New York (2011)
17. Chromik, M., Eiband, M., Buchner, F., Krüger, A., Butz, A.: I think I get your point, AI! The illusion of explanatory depth in explainable AI. In: Proceedings of the 26th International Conference on Intelligent User Interfaces, pp. 307–317. ACM (2021)

18. Bove, C., Laugel, T., Lesot, M.-J., Tijus, C., Detyniecki, M.: Why do explanations fail? A typology and discussion on failures in XAI. arXiv preprint arXiv:2405.13474 (2024)
19. Rozenblit, L., Keil, F.: The misunderstood limits of folk science: An illusion of explanatory depth. Cogn. Sci. **26**(5), 521–562 (2002)
20. Schemmer, M., Kühl, N., Benz, C., Bartos, A., Satzger, G.: Appropriate reliance on AI advice: Conceptualization and the effect of explanations. In: Proceedings of the 28th International Conference on Intelligent User Interfaces (IUI '23), pp. 410–422. ACM (2023).
21. Nickerson, R.S.: Confirmation bias: A ubiquitous phenomenon in many guises. Rev. Gen. Psychol. **2**(2), 175–220 (1998)
22. Wexler, J., Pushkarna, M., Bolukbasi, T., et al.: The what-if tool: Interactive probing of machine learning models. IEEE Trans. Vis. Comput. Graph. (2019)
23. Carroll, J.M., Rosson, M.B.: Paradox of the active user. In: Carroll, J.M. (ed.) Interfacing Thought: Cognitive Aspects of Human-Computer Interaction, pp. 80–111. MIT Press, Cambridge (1987)
24. Wall, E., Blaha, L.M., Franklin, L., Endert, A.: Warning, bias may occur: A proposed approach to detecting cognitive bias in interactive visual analytics. In: IEEE Conference on Visual Analytics Science and Technology (VAST), pp. 104–115. IEEE, Phoenix (2017)
25. Lundberg, S.M., Lee, S.-I.: A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems, vol. 30, pp. 4765–4774 (2017)