Computers in Human Behavior Reports 22 (2026) 101050

Contents lists available at ScienceDirect

Computers in Human Behavior Reports

journal homepage: www.sciencedirect.com/journal/computers-in-human-behavior-reports

Counterfactual explanations with varying parameter control: Effects on mental-model formation and explanation satisfaction

Lena Kölmel [a,*], Maximilian Becker [b], Finn Schwall [c], Jutta Hild [d], Pascal Birnstill [d], Barbara Deml [a], Jürgen Beyerer [b]

[a] Institute of Human and Industrial Engineering (ifab), Department of Mechanical Engineering, Karlsruhe Institute of Technology, Karlsruhe, Germany
[b] Vision and Fusion Laboratory (IES), Department of Informatics, Karlsruhe Institute of Technology, Karlsruhe, Germany
[c] Simula, Oslo, Norway
[d] Fraunhofer Institute of Optronics, System Technologies and Image Exploitation (IOSB), Karlsruhe, Germany

ARTICLE INFO

Keywords:
Human-AI-Interaction
Explainable artificial intelligence
Interactive explanations
Counterfactuals
User study

ABSTRACT

Algorithms increasingly inform consequential decisions, yet their workings are often opaque. Interactive explanation interfaces are often assumed to enhance comprehension, but prior comparisons to static baselines frequently confound user agency with unequal informational content. We test whether greater agency over explanation-generation parameters improves understanding when informational parity is ensured. In a between-subjects laboratory study (N = 64), non-experts worked with counterfactual explanations of a binary classifier using one of three interface configurations: a fixed condition with no parameter control, a system-randomized condition in which parameter settings varied automatically across generations, or a user-controlled condition in which participants adjusted feature weights and exclusions. All conditions provided equivalent informational content and allowed repeated re-generation of counterfactual sets; only the locus of control over generation parameters differed. Behavioral logs captured exploratory breadth, in-depth exploration, exclusion rate, and interaction time per data point. Outcome measures included objective understanding, self-reported understanding and confidence, explanation satisfaction, and cognitive workload. Parameter control increased exploratory breadth relative to the fixed condition. The system-randomized condition yielded the greatest in-depth exploration and the longest interaction time per data point, whereas the user-controlled condition produced higher exclusion rates. Despite these behavioral differences, objective understanding and explanation satisfaction did not differ between conditions; self-reported understanding was highest in the fixed condition. Mental demand and frustration were highest in the user-controlled condition. Overall, varying the locus of control over generation parameters primarily changed how and how much participants explored counterfactual explanations without improving objective understanding or satisfaction under informational parity, while increasing subjective workload when control was user-driven.

1. Introduction

AI-based and algorithmic systems increasingly support human decision-making across a wide range of domains (for an overview, see Islam et al., 2022; Ullah et al., 2025). This growing prevalence has led to calls from research, industry, and policymakers that users have a right to transparency and accountability in decisions made with or by automated algorithmic systems. Article 22 of the European Union General Data Protection Regulation (GDPR) mandates that data subjects receive meaningful information about the logic involved in automated decision-making systems (EU, 2016). The field of explainable artificial intelligence (XAI) aims to render opaque AI behavior more transparent through various algorithmic methods (for an overview, see Dwivedi et al., 2023). Miller et al. (2017) caution that "the inmates are running the asylum," highlighting that XAI approaches are often developed by data science experts for other data science experts, lacking the synergetic integration of cognitive and social sciences necessary to enhance intelligibility for non-expert users.

* Corresponding author. Institute of Human and Industrial Engineering (ifab), Department of Mechanical Engineering, Karlsruhe Institute of Technology, Engler-Bunte-Ring 4, 76131, Karlsruhe, Germany.
E-mail address: lena.koelmel@kit.edu (L. Kölmel).

https://doi.org/10.1016/j.chbr.2026.101050
Received 25 November 2025; Received in revised form 26 March 2026; Accepted 28 March 2026
Available online 1 April 2026

Recent research on human-centered explanations has advocated for the design of interactive AI explanations to overcome the intelligibility limitations of XAI methods and to enhance, in particular, end-user acceptance of AI-based systems (Arya et al., 2019; Langer et al., 2021; Raees et al., 2024). Interactive explanations enable AI behavior and decisions to be communicated in ways that resemble social, dialogue-based explanatory processes between humans, wherein an explainer provides targeted information to an explainee, accompanied by opportunities to ask clarifying follow-up questions (Miller et al., 2017; Weld & Bansal, 2019). Prior work suggests that humans expect and prefer explanations of complex systems to be delivered in this manner (Yi et al., 2007).

Empirical studies report mixed effects of interactive explanations on different aspects of human-AI-interaction like for example, perceived usefulness, team performance, and understanding, or cognitive load and overreliance (Bertrand et al., 2023). Most prior work either evaluates a single interactive interface without a control condition (Cheng et al., 2022; Das Antar et al., 2024; Hernandez-Bocanegra & Ziegler, 2021) or compares a static interface to an interactive one (Bove et al., 2022; Liu et al., 2021). However, in many of these studies, an informational inequality exists between the static and interactive conditions: through interactive manipulation of explanations, users may acquire more knowledge about the algorithm, decision rules, and underlying machine learning model. Consequently, the reported positive effects on objective understanding and mental model formation may partly result from the additional information available in the interactive condition. To our knowledge, no empirical study has systematically compared different modes of interactivity while simultaneously controlling for informational equality between conditions. This study addresses that gap by comparing different interface configurations under informational equality, while making a critical distinction between two interaction mechanisms. Building on Bertrand et al.'s (2023) taxonomy, we conceptualize interactivity primarily as agency over "mutate" actions, i.e., reconfiguring counterfactual generation parameters through feature weighting and/or feature exclusion, whereas iterative exploration via repeated re-generation of counterfactual sets is treated as a shared exploration affordance. Accordingly, iterative re-generation was available in all conditions to ensure informational equality; conditions differed only in who controlled the mutate actions: parameters were held constant in a fixed condition, a system-randomized condition, and a user-controlled condition.

We focus on counterfactual explanations as the XAI method under investigation. This method explains the prediction of a black-box machine learning model by providing a counterfactual example, defined as the minimal changes that must be applied to a single data instance to obtain a different (often more favorable) prediction. Counterfactuals follow a what-if logic, illustrating why the model did not produce a certain prediction (Guidotti, 2024). Much like a teacher justifying a mid-level grade by specifying what would have been required for a higher grade, this contrastive reasoning is intuitively understandable, offers explanatory value, and provides a causal dimension that distinguishes counterfactuals from other XAI methods (Adadi & Berrada, 2018; Warren et al., 2022). Counterfactuals are therefore particularly suitable for examining the effect of interactive explanations on the formation of adequate mental models and explanation satisfaction in non-expert users. We focus on non-expert users because prior research shows that they generally struggle most with understanding how AI-based systems work and how model input defines model output (Cheng et al., 2019; Tullio et al., 2007).

This paper makes three primary contributions to the literature on human–AI interaction and XAI. First, we provide, to our knowledge, the first empirical comparison of explanation interfaces that vary in interactivity as user agency over "mutate" actions (feature weighting and exclusion), while holding iterative exploration via repeated re-generation constant across conditions. This extends prior work by clarifying whether increased interactivity and user control over generation parameters are inherently beneficial or whether they risk cognitive overload, particularly for non-expert users. Second, we isolate the mechanism underlying observed benefits of parameter-reconfigurable explanation interfaces. By ensuring informational equality across the two non-static interfaces (system-randomized versus user-controlled parameterization), we disentangle whether positive effects arise from active user control over mutate actions or from informational advantages that can accompany richer interfaces. Third, we advance research on counterfactual explanations by directly comparing a fixed-parameter baseline, a system-randomized-parameter interface, and a user-controlled-parameter interface with respect to perceived user-friendliness and understanding, testing whether - and under which form of parameter control - counterfactual explanations offer measurable advantages over a static baseline.

2. Related work and research questions

2.1. Interactive XAI

There is currently no unified definition of interactivity in the context of human-(X)AI-interaction. Given the field's interdisciplinary orientation, arriving at a single definition that integrates all disciplinary perspectives remains challenging. Moreover, different disciplines often rely on different terminology. Across the research landscape, the terms adaptive (Turchi et al., 2024), personalized (Silva et al., 2024), and customized (Esposito et al., 2025) frequently appear in the context of XAI and can be subsumed under the umbrella term interactivity. Early work conceptualized interactivity as communication between humans and a computer or complex system (Dix & Ellis, 1998; Foley et al., 1996). More recent work defines interactivity as the control and action between a human and an artifact or system (Janlert & Stolterman, 2017). Both understandings readily apply to the XAI context: compared with static explanations, interactive explanations respond to Miller et al.'s (2017) call for an interactive explanation process by enabling users' direct influence on the explanatory process.

It is also essential to delineate the phase of the system lifecycle to which interactivity pertains. Interactive machine learning (iML), originating with Fails and Olsen (2003), involves the interactive inclusion of users, typically data scientists or other experts, during the training phase. This approach commonly aims to improve model performance when data is scarce or of insufficient structure or quality. Through a train–feedback–correct loop, users take the role of co-trainers and remain in the loop of the development process (Gómez-Carmona et al., 2024). By contrast, interactive AI can span the entire lifecycle of a system. In interactive designs, the focus is on continuous, reciprocal collaboration and on improving human–AI teaming (Raees et al., 2024). Depending on the implementation, interactive XAI is closely related to iML but targets end users and other typically non-expert stakeholders. Its aim is to enable users to customize the computation or the presentation of explanations (or both), thereby increasing model transparency and, in turn, trust, willingness to use, and users' mental models of the system (Bertrand et al., 2023). The present study focuses on interactive XAI. However, because the experimental task is designed such that participants attend exclusively to the explanations and the underlying machine learning model (see Section 3.1 for details) without the integration of a decision-making scenario, the distinction from iML is blurred.

Psychological and educational research offer several theoretical arguments for the benefits of interactivity. From a constructivist perspective, "learning by doing" is central to knowledge acquisition (Roussou, 2004; Skulmowski, 2024), as people learn by simulating and testing hypotheses about personally meaningful real-world relationships. In accordance with this, Evans and Gibbons (2007) found that interactive multimedia interfaces promote deeper learning than static ones by actively engaging learners. In the context of XAI, Cabrera et al. (2023) analyzed the cognitive sensemaking processes of data scientists and how they develop rich mental models of AI behavior. Their

framework describes a continuous and iterative cycle of data scientists identifying new instances, revising hypotheses, and updating their mental models, which also advocates the necessity of interactivity for AI related sensemaking.

Empirical results do not always align with these theoretical assumptions. Overall, based on the existing literature, it remains difficult to draw empirically well-supported, unambiguous conclusions about the mechanisms by which interactivity affects different aspects of human–AI interaction. The few experimental studies available are heterogeneous in their research focus, the XAI methods examined, and the operationalization of interactivity, limiting both comparability and generalizability (Bertrand et al., 2023). For example, Cheng et al. (2019) compared static with interactive explanations in a student-application scenario and reported significant effects of interactivity on both subjective and objective understanding, although interactivity was accompanied by increased processing time. In their study, explanations could be modified with sliders that assigned weights to local feature-importance–based visual explanations. By contrast, Bove et al. (2022) did not find improvements in objective understanding due to an interactive interface, although participants reported higher satisfaction with the interactive explanation design; here, interactivity was limited to a sorted ranking and display of relevant features in an insurance scenario. A similar pattern emerged in a study that examined the effects of different XAI methods, presented in both static and interactive formats, on human–AI team performance: with interactive explanations, participants did not outperform the standalone AI, yet they rated the AI assistance as more helpful for some models (Liu et al., 2021). In a gamified setting, Holzinger et al. (2019) could observe positive effects on human-AI team performance, by letting participants directly control an ant-colony algorithm. These mixed results highlight that the effectiveness of interactivity is highly context-dependent and further research is needed.

2.2. Interactive XAI and cognitive workload

There has been growing concern in the human-computer-interaction (HCI) community that overly complex XAI techniques and their mode of presentation increase cognitive workload (Hudon et al., 2021; Koh et al., 2025; Liao et al., 2020). Especially for non-expert users with no prior exposure to XAI methods, the presentation of statistical metrics, plots, and tables often appears inaccessible and overwhelming (Liao et al., 2020; Xie et al., 2020). This is not merely a usability issue, as high cognitive workload has been associated with negative effects of explanations, such as overreliance (Wang & Yin, 2021; Zhang et al., 2020). Interactive explanations introduce an additional dimension of complexity that can further overwhelm users. Based primarily on qualitative data, Panigrahi et al. (2025) report that some interactivity mechanisms were, at times, overwhelming. Inexperienced users may struggle to correctly understand the available interaction options in addition to the explanations themselves and, consequently, to use them productively. Moreover, a portion of available cognitive resources must be invested in understanding how the interactivity mechanisms function, thereby diverting cognitive engagement from the task itself. This could result in less rich mental models about the AI system with an emphasis on the formation of functional mental models (‘how to use’) rather than structural mental models (‘how and why it works’) (Johnson-Laird, 1983). Constraining interaction may help reduce users' cognitive workload while still outperforming static interfaces. For example, Dietvorst et al. (2018) found that users' preference for modifiable algorithms was indicative of a desire for some control over the algorithm's workings rather than for greater control. Consistent with this, a literature review grounded in a design science research approach recommends increasing user agency, for example, through unrestricted interactivity, while simultaneously decreasing complexity and cognitive effort (Speckmann et al., 2025). Similarly, Kulesza et al. (2015) formulate design principles demanding sound and complete but also not overwhelming interactive explanations.

2.3. Counterfactual explanations in interactive XAI

There are various existing methods to generate counterfactual explanations. These methods differ with regard to how they define the minimal changes between the referenced instance and the generated counterfactuals (Guidotti, 2024). Prior work has focused on integrating constraints into these definitions of minimal distance to generate counterfactuals that are feasible (Poyiadzi et al., 2020), plural and diverse (Bove et al., 2023; Mothilal et al., 2020), plausible (Kenny & Keane, 2021; Warren et al., 2022), or that account for users’ cognitive processes (Celar & Byrne, 2023; Jeyasothy et al., 2022). Although recent work emphasizes that user studies testing counterfactual-based XAI methods remain sparse, as many papers primarily focus on method development and optimization (Shang et al., 2022; Verma et al., 2024), it has been demonstrated that counterfactuals elicit causal reasoning in users (Byrne, 2019). This makes them potentially useful for fostering the development of rich mental models about AI systems. A recent review on explainable user interfaces (XUIs), defined as the sum of outputs from an XAI system with which the user can directly interact (Chromik & Butz, 2021), found that counterfactuals are the second most frequently used XAI method in this research domain (Cappuccio et al., 2025). Similar to research on XAI methods, most XUIs have not been evaluated in controlled user studies, though some prior work links interactive explanations and counterfactuals. For example, Hohman et al. (2019) developed an interactive user interface for providing explanations to data scientists and reported that these expert users most frequently employed counterfactuals for comparing and exploring model predictions. ViCE (Gomez et al., 2020) and AdViCE (Gomez et al., 2021) are two prototype implementations for visual interactive counterfactuals, whose functionality was demonstrated in case studies rather than controlled user studies. An interface based on interactive counterfactuals, designed to support understanding of large language model behavior, was rated as both useful and enjoyable to use by experts and non-experts alike (Cheng et al., 2024). Similarly, Myers et al. (2020) presented an interactive counterfactual visualization for AI non-experts to explore the decisions of a semantic neural network in the context of loan applications, aiming to reveal potential bias. An expert panel comprising AI, HCI, and social science scholars evaluated the application as useful in reducing cognitive biases. Bove et al. (2023) reported positive effects on both objective and subjective understanding, as well as explanation satisfaction, when participants explored plural rather than single counterfactuals in a loan application scenario. Suffian et al. (2025) found that counterfactuals with interactive user feedback outperformed counterfactuals without feedback in objective understanding, satisfaction, and trust.

2.4. Research questions and hypotheses

This paper investigates how different modes of parameter control in a counterfactual explanation interface affect non-expert users. We distinguish iterative exploration through re-generating counterfactual sets from user agency over “mutate” actions, namely feature weighting and feature exclusion. First, we examine whether exploration behavior, specifically the depth and breadth of exploration, differs across conditions. Second, we analyze effects on the formation of a rich and adequate mental model and on explanation satisfaction. In contrast to prior work, we compare a fixed-parameter baseline with two non-fixed configurations that differ in the locus of control over mutate actions, namely system-randomized versus user-controlled parameterization. The option to re-generate counterfactual sets was available in all conditions to ensure informational equality. We expect system-randomized parameterization to be less overwhelming for non-expert users without prior experience in interpreting counterfactuals, as it reduces the operational burden of actively configuring interaction settings. Productive use of a

user-controlled interface may presuppose that users hold hypotheses, for example about relationships among features in the data model or about the algorithm's performance limits, that they can test by adjusting generation parameters. We therefore test whether a less demanding, system-randomized interface better supports non-expert users than an interface requiring active user control over mutate actions, under informational equality. More precisely, the study is guided by the following research questions and hypotheses:

- **RQ1:** To what extent does the mode of parameter control in the explanation interface (fixed, system-randomized, user-controlled) affect non-experts' mental model adequacy and explanation satisfaction when interacting with counterfactual explanations?
  - **H1.1**: Users' objective understanding is highest in the system-randomized condition.
  - **H1.2**: Users' self-reported understanding is highest in the system-randomized condition.
  - **H1.3**: Users are most confident in their understanding of the algorithm in the system-randomized condition.
  - **H1.4**: Users' satisfaction with the counterfactual explanations is highest in the system-randomized condition.
- **RQ2**: How is cognitive workload during counterfactual exploration affected by the mode of parameter control (fixed, system-randomized, user-controlled)?
  - **H2**: Users have the highest cognitive workload in the user-controlled condition.
- **RQ3:** How does exploration behavior differ across interface configurations with different modes of parameter control (fixed, system-randomized, user-controlled)?

We address this research question through exploratory analyses of behavioral log data. Behavioral log data were used to derive parameters capturing multiple facets of exploratory behavior. Because these parameters are specific to the experimental application used in this study, and because there is limited prior empirical work describing exploration behavior at this level of granularity, we refrain from formulating specific hypotheses.

3. Method

3.1. Task description: an abstract and binary classification scenario

During the study, participants interacted with a browser-based experimental application that offered basic functionalities such as selecting data points and generating counterfactual explanations. The participants' task was to identify dependencies between features in the underlying data set by analyzing ten available data points and the corresponding explanations. Each data point consisted of six features presented in an abstract format, labeled alphabetically from A to F. No accompanying scenario was provided to contextualize or semantically enrich the meaning of the features. Presentation was limited to these alphabetical labels, and the dataset was artificially generated (see Section 3.4).

The general setup of the task aligns with scenarios commonly used in human-(X)AI-interaction research. Many studies employ real-world decision-making contexts such as human resources, finance, insurance, medicine, or law (Islam et al., 2022; Kalasampath et al., 2025). Especially when working with layperson samples, more accessible domains such as music preferences (Martijn et al., 2022) or food choices (Buçinca et al., 2021) are used, as they require no specific domain knowledge and there is less bias due to assumptions about real world dependencies. Scenarios that simulate decision support systems in professional domains are particularly valued for their high external validity, allowing results to be generalized to real-world contexts with comparable complexity. However, world knowledge-based assumptions about dependencies between features can obscure the isolated effects of AI explanations and interaction modes on human-AI-interaction. When an algorithm aligns with common pre-existing beliefs, it becomes difficult to determine whether participants genuinely understand the underlying model or whether the algorithm's behavior merely conforms to their expectations. Conversely, when the algorithm deviates from those beliefs, it may induce biased perceptions of feature-outcome relationships and lead to negative evaluations of the AI system. Cheng et al. (2019) addressed this challenge by introducing 'unnamed attributes' to assess objective understanding of algorithmic reasoning, independent of domain knowledge. Building on this idea, we deliberately employed an abstract scenario to minimize prior assumptions and isolate comprehension of the model logic.

3.2. Data set

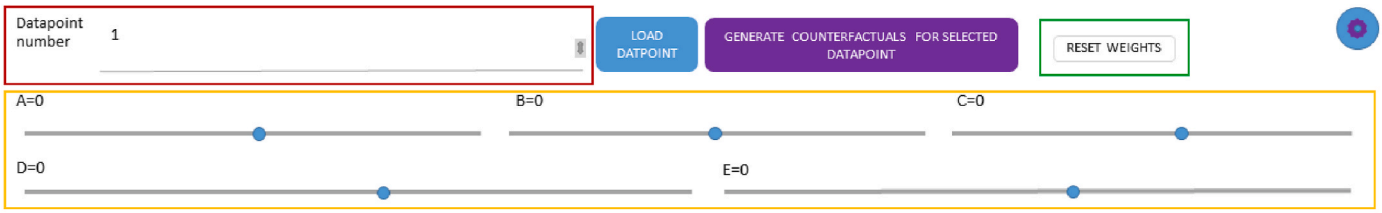
The application was based on a synthetically generated dataset consisting of six features, labeled A through F. Feature F represents the binary classification outcome, while features A to E serve as potential predictors. The predictor features span distinct numerical ranges: A [0–100], B [0–20], C [0–50], D [0–30], and E [0–90]. To minimize cognitive load and memory demands associated with the abstract, label-only representation, the number of features was deliberately limited. The dependencies between features are defined such that A, B, and C influence the outcome F. In contrast, D and E are interdependent but do not affect the outcome. The dataset is constructed by randomly sampling from a uniform distribution for features A, B, C, and D. Subsequently, features E and F are computed by combining the previously sampled values using additive and subtractive operations, with additional Gaussian noise. In total, 10,000 data points were generated using this procedure.

3.3. Interactive application

The application used in this study was implemented in RIXA (Becker et al., 2023), a modular platform designed primarily for explainable artificial intelligence[1]. RIXA integrates a natural language chat interface with a visual dashboard and supports the integration of custom XAI tools through a plugin system. While the chat functionality was not utilized in the present study, RIXA was selected due to its built-in support for XAI-relevant components. A customized implementation of DiCE (Mothilal et al., 2020) was integrated as a plugin to generate counterfactual explanations. For the underlying model, a random forest classifier implemented via scikit-learn (Pedregosa et al., 2011) was employed. The user interface of the application (illustrated in Fig. 1) comprises three main functionalities:

- **Data Point Selection.** Ten pre-selected data points were available via a drop-down menu. These were selected from different regions of the feature space to ensure a diverse set of counterfactual explanations and facilitate participants' understanding of feature dependencies. All data points share the same predicted class (zero) and remain constant across all experimental conditions and participants. Once selected, a data point is presented in a single-row table with features labeled A to F as columns.
- **Counterfactual Generation.** Upon clicking the corresponding button, ten counterfactuals are generated and displayed beneath the selected data point. Unchanged features are marked with a dash. Participants may request counterfactuals repeatedly, but only ten are shown at any given time.
- **Customization of Counterfactuals.** Sliders are provided for features A to E to specify the parameterization of counterfactual generation via feature weighting and feature exclusion. By default, all

[1] The source code of the application used in this study is available on GitHub: https://github.com/counterfactual_user_study.



A	B	C	D	E	F
44.0	8.0	32.0	16.0	42.0	0
48.0	-	-	-	-	1
-	-	34.0	-	-	1
49.0	-	-	-	-	1
47.0	-	33.0	-	-	1
45.0	-	34.0	-	-	1
48.0	-	-	-	43.0	1
48.0	-	-	-	41.0	1
-	-	34.0	-	43.0	1
-	-	34.0	-	41.0	1
50.0	-	-	-	-	1

Fig. 1. Counterfactual explanation user interface used in the study. The interface comprises five main components: (a) a drop-down menu for selecting data points (red), (b) a button to generate counterfactuals (purple), (c) sliders to adjust parameters for counterfactual generation (orange), (d) a "reset weights" button that restores weights to their neutral default settings (green), and (e) a table displaying the feature values of the selected data point (top row, gray background) alongside ten generated counterfactuals (subsequent rows). In the fixed condition, the slider panel (orange) and the reset button (green) are disabled. In the system-randomized condition, the slider panel is locked (read only); parameter settings are assigned automatically for each generated counterfactual set and cannot be adjusted manually. In the user-controlled condition, slider values can be adjusted prior to generation and reset to the default configuration. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

slider values are centered at zero and the interface provides a reset function that restores this default configuration. Higher slider values increase the corresponding feature weight, reducing the likelihood that the feature is altered in generated counterfactuals. Lower slider values decrease the weight, making the feature more likely to vary. The slider range spans from -5 to $+5$, and setting a slider to -5 excludes the feature from counterfactual generation. If the parameter configuration is too restrictive to allow valid counterfactuals, an error message is shown. The availability of parameter control and reset functionality varied by experimental condition and is described in Section 3.4.

3.4. Experimental conditions

We implemented three experimental conditions that differed in the locus of control over mutate actions, namely feature weighting and feature exclusion. In all conditions, participants could select from the predefined data points and request counterfactual sets repeatedly; conditions differed only in whether generation parameters were fixed, assigned by the system, or controlled by participants.

- **Fixed condition.** In this condition, participants had no access to parameter control. All feature weights were fixed at 0 and feature exclusion was not available. The slider panel and the reset functionality were disabled. Participants could freely select data points and request an unlimited number of counterfactual sets, which were generated under the same fixed parameterization.
- **System-randomized condition.** In this condition, participants could freely select data points and request an unlimited number of counterfactual sets. However, parameter control was not available. Upon each generation request, slider values were assigned automatically by the system and could not be adjusted manually; the slider panel was displayed but locked. The randomization was defined such that, for each feature, there was a 15% probability of exclusion (slider value set to -5), and all other values were sampled at random within the range -5 to $+5$.
- **User-controlled condition.** In this condition, participants had full access to parameter control. In addition to selecting data points and requesting counterfactual sets, they could adjust feature weights via sliders and exclude features by setting a slider value to -5. Participants could also reset all parameters to the default (0) configuration at any time.

3.5. Calculation of counterfactuals

We implemented an extended counterfactual generation procedure based on DiCE (Mothilal et al., 2020) that supports feature weighting and feature exclusion during the search. To keep the study focused, we deliberately disabled DiCE options that are orthogonal to our research question (e.g., explicit diversity terms, plausibility constraints). The objective is to produce a set of counterfactuals that (i) achieve a desired model prediction and (ii) stay as close as possible to the original instance.

3.5.1. Core mechanism

We employ an evolutionary search:

- **Population-based search**: A population of candidate counterfactuals is maintained across iterations.
- **Mutation and mating**: Candidates are perturbed by (i) random changes within valid ranges (mutation) and (ii) exchanging feature values between candidates (crossover).

- **Selection**: Candidates are scored by a loss function; the best are retained for the next generation.

3.5.2. Loss function

The algorithm optimizes for two primary components:

- **Target outcome loss**: This component ensures the counterfactual achieves the desired prediction outcome. It is a binary penalty that evaluates to 0 when the prediction is changed and 100 otherwise. Either the candidate meets the target criterion or incurs a high penalty.
- **L1 distance loss**: Minimizes the Manhattan distance between the counterfactual and the original instance. This ensures the counterfactual represents minimal changes:

$$L_1 = \sum_i w_i \frac{x_i - x_i^{original}}{span_i}$$

Where:

- w_i represents feature weights as defined by the interface configuration (fixed condition, system-randomized condition, user-controlled condition).
- Features are normalized by their ranges ($span_i$).
- The total loss is the sum of these components.

3.5.3. Interface-level weight transformation

Weights between [0, 1] increase the importance (as the loss decreases), while values above 1 decrease the importance of a feature. The impact of the weights is exponential. To reduce mental load in the interface, slider values are transformed using a polynomial into a linear scale between −5 and 5, where 0 corresponds to an actual weight of 1, -5 excludes a feature, and +5 corresponds to a weight close to 0.

3.5.4. Termination

The search terminates when (i) the desired number of counterfactuals has been found, (ii) the improvement in the loss over several consecutive iterations falls below a predefined tolerance, or (iii) the maximum number of iterations is reached. If the search terminates before a sufficient number of counterfactuals has been generated, an error message indicating overly restrictive parameter settings is displayed in the interface.

3.5.5. Counterfactual selection

The counterfactuals presented in the interface are those with the lowest total loss after the algorithm terminates. This means the interface presents instances that (1) successfully change the prediction to the desired outcome and (2) require minimal changes to the original instance (measured by L1 distance).

3.6. Materials

Various methods and evaluation criteria have been proposed to assess the adequacy and effectiveness of AI explanations from a human-centered perspective (for an overview see Hoffman et al., 2023; Miller, 2019). There is broad consensus that both objective understanding and subjective satisfaction must be considered to obtain a holistic picture of human–AI-interaction (see also Naveed et al., 2024; Silva et al., 2023). In line with this, we evaluate the effectiveness of the proposed explainable user interfaces using both types of measures.

3.6.1. Mental model and confidence

Our approach to measuring objective understanding follows work by Cheng et al. (2019) and is grounded in the conceptualization by Weld and Bansal (2019), who define human understanding of an algorithm as the ability to identify which features drive its decisions and to anticipate how changes in input would influence the output. Based on this framework, we developed three types of questions to assess participants' objective understanding of the counterfactual explanations.

- **Feature Dependencies**: In line with the primary goal of the experimental task, identifying dependencies between features in the dataset, participants were asked to indicate which features directly influence the outcome variable F.
- **Decision Prediction**: To assess participants' holistic understanding of the data model and the counterfactual explanations, participants were presented with new, previously unseen data points and asked to predict the binary classification outcome. Each question offered two response options. Three items of this type were included in the final questionnaire.
- **Alternative Prediction**: To evaluate participants' ability to anticipate which changes would reverse the classification outcome, participants were shown a new data point and asked which feature change would most likely lead to the opposite classification. Each item offered three response options. Two items of this type were included in the questionnaire.
- **Objective Understanding**: We compute a mean across these three indicators to operationalize overall objective understanding. Internal consistency was acceptable, $\alpha = .71$.

To ensure a comprehensive operationalization, the objective measures were supplemented with the following self-report-based variables.

- **Confidence**: To capture participants' level of certainty, they also rated their confidence in their answer regarding the feature dependencies using a 7-point Likert scale.
- **Self-reported understanding**: We additionally measured the self-reported understanding of participants with a single item ("*I understand how the application's algorithm works*"), using a 7-point Likert agreement scale.

3.6.2. Explanation satisfaction

To assess participants' satisfaction with the counterfactual explanations, we adapted the Explanation Satisfaction Scale (Hoffman et al., 2023). The original scale comprises eight items; two items were excluded due to their limited applicability to the abstract scenario used in the present study. Participants rated their agreement with each statement on a 7-point Likert scale. Internal consistency for the overall questionnaire was acceptable, $\alpha = .77$.

3.6.3. Cognitive workload

To assess cognitive workload during interaction with the application, the *Mental Demand* and *Frustration* subscales of the German translation of the *NASA Task Load Index* (NASA-TLX; Hart, 2006) were used.

3.6.4. Exploratory behavior

Based on the behavioral data obtained from participants' interaction with the application, four parameters were extracted. The parameters represent frequency-based indices, including data point selection, counterfactual generation, and the occurrence of feature exclusion during generation. Conceptually, we subsume these measures under the umbrella term exploratory behavior.

- **Exploration score**: Frequency-based aggregate representing the overall extent of exploratory activity by combining the number of selected data points and the number of generated counterfactuals into a single sum score. It reflects the total amount of exploratory activity, independent of the specific interface condition, and thus provides a common baseline for comparing exploratory behavior across all experimental groups.
- **Relational exploration score**: As a complementary measure, this score reflects the ratio of generated counterfactuals to selected data

points (lower values indicate more counterfactuals sets generated per data point). It captures the degree to which participants engaged in in-depth exploration by generating multiple explanations per data point.

- **Exclusion rate**: Proportion of feature-settings excluded during counterfactual generation (slider value = −5), normalized by the number of exclusion opportunities. Specifically, for each participant we computed the number of excluded feature-settings divided by the total number of feature-settings presented across all generation requests (CounterfactualCounter × 5 features A–E). In the user-controlled condition, the rate reflects participants' parameter choices. In the system-randomized condition, the rate reflects system-assigned parameterization (expected rate 0.15 by design, with sampling variability). Exclusion was not available in the fixed condition.
- **Interaction time per data point**: The time spent interacting per data point, measured in seconds.

3.7. Procedure

The study was conducted as an in-person laboratory experiment. Upon arrival, participants reviewed an informed consent form and detailed written instructions, followed by a structured introduction to the experimental task. An online tutorial explained the task, summarized the computational principles underlying counterfactual explanations, and introduced the functionality of the experimental application. Information about parameter control options for counterfactual generation varied by assigned experimental condition (fixed, system-randomized, user-controlled), while all other tutorial content was held constant across conditions. Participants were randomly assigned to one of the three experimental conditions before starting the tutorial. To ensure adequate understanding, participants then completed four single-choice comprehension questions; those who answered any item incorrectly were excluded from the analytic sample prior to hypothesis testing. After successfully completing the tutorial, participants engaged in a 30-min interaction phase during which they could freely explore the application and the preselected data points. Finally, they completed an online questionnaire assessing the outcome measures and demographic information.

3.8. Participants

We recruited students via local advertisements on campus and online. To ensure a comparable level of familiarity with AI related topics, we defined a technical or engineering major as an inclusion criterion. Six participants were excluded because they did not pass the attention check, which consisted of three or four, depending on the experimental condition, single-choice questions about the definition of counterfactuals and the functional features of the application. Participation was remunerated with 12 € for an estimated experiment duration of 1 h. The final sample consisted of 64 participants aged 19 to 40 ($M = 23.2$, $SD = 3.57$). Approximately two-thirds were male ($n = 41$), and about half ($n = 30$) were currently studying computer science at the local university in Karlsruhe, Germany. Participants reported intermediate prior knowledge of AI ($n = 29$ indicated some prior knowledge, and n = 16 reported profound prior knowledge), but little prior knowledge of XAI ($n = 40$ indicated no prior knowledge, and $n = 21$ reported little prior knowledge). No significant differences in any aspect of prior knowledge were observed across the three experimental conditions based on the results of a one-way ANOVA ($F(2, 61) = 0.28$, $p = .755$ for prior knowledge about AI and $F(2, 61) = 0.06$, $p = .941$ for prior knowledge about XAI). Five participants were familiar with the term counterfactual explanations prior to the study. This was not an exclusion criterion, as a tutorial preceded the interaction phase to align participants' knowledge levels. The Ethics Committee of the Karlsruhe Institute of Technology (KIT) issued a certificate of non-objection for this study.

3.9. Statistical analyses

One-way analyses of variance (ANOVA) were conducted to examine group differences among the three experimental conditions (fixed, system-randomized, user-controlled). Post hoc comparisons were performed using Tukey's honestly significant difference (HSD) test. For measures that were only available in the system-randomized and user-controlled conditions (e.g., exclusion), we conducted planned pairwise comparisons using independent-samples t-tests; when assumptions were violated, Mann–Whitney U tests were used. The alpha level was set at 0.05. All analyses were conducted using SPSS Statistics for Windows, Version 27.0 (IBM Corp, 2020).

4. Results

We present results starting with exploratory behavior (RQ3). To align the presentation with the locus-of-control framing, we highlight differences between the two non-user-controlled interfaces (fixed and system-randomized) and the user-controlled interface within each results section. Descriptive statistics are reported in Table 1.

4.1. Exploratory behavior

To address RQ3, we first examined whether exploratory behavior differed as a function of condition (fixed, system-randomized, user-controlled). To compare overall exploratory activity across all three conditions, we used the exploration score as a common baseline (Section 3.6.4). We found a significant effect of condition on exploratory behavior, $F(2, 61) = 8.15$, $p < .001$, $\eta^2 = 0.21$. Post hoc comparisons indicated that participants in the user-controlled condition engaged in more exploratory activity ($M = 124.14$, $SD = 52.75$) than those in the fixed condition ($M = 62.55$, $SD = 27.64$).

We additionally analyzed whether in-depth exploration differed between conditions using the relational exploration score, calculated as the ratio of the number of selected data points to the number of generated counterfactual sets, such that lower values indicate that participants generated more counterfactual sets per data point. This analysis also revealed a significant effect of condition, $F(2, 61) = 7.95$, $p < .001$, $\eta^2 = 0.34$. Descriptively, relational exploration was lowest in the system-

Table 1
Descriptive statistics by condition.

Measure	fixed		system-randomized		user-controlled	
	M	*SD*	*M*	*SD*	*M*	*SD*
Explanation satisfaction	4.60	1.13	4.79	0.95	4.88	1.07
Exploratory behavior						
Exploration score	62.55	27.64	98.62	64.27	124.14	52.75
Relational exploration score	1.00	0.30	0.37	0.35	0.64	0.46
Exclusion rate*			0.14	0.02	0.33	0.29
Interaction time per data point	64.71	35.33	114.09	79.42	51.79	21.53
Mental Model						
Feature dependencies	3.86	0.71	3.95	0.87	3.52	1.03
Decision prediction	1.41	0.80	1.38	0.87	1.57	0.68
Alternative prediction	2.45	0.67	2.33	0.86	2.33	0.80
Objective understanding	7.73	1.36	7.67	1.60	7.43	1.66
Self-reported understanding	4.27	1.78	3.52	1.44	3.05	1.50
Confidence	4.55	1.95	4.19	1.40	4.00	1.41
Cognitive Workload						
Frustration	9.27	3.04	7.90	2.74	12.76	3.03
Mental Demand	10.09	2.24	12.43	3.60	14.61	3.56

Note. *Exclusion rate not applicable in fixed condition.

randomized condition (M = 0.37, SD = 0.35), followed by the user-controlled condition (M = 0.64, SD = 0.46), whereas the fixed condition was close to one (M = 1.00, SD = 0.30), indicating approximately one generated set per selected data point.

For feature exclusion, we compared exclusion rates between the system-randomized and user-controlled conditions (exclusion not available in the fixed condition). A Mann–Whitney *U* test indicated higher exclusion rates in the user-controlled condition (M = 0.33, SD = 0.29) than in the system-randomized condition (M = 0.14, SD = 0.02; expected rate 0.15 by design), U = 136.00, z = −2.13, p = .034, r = 0.33.

Finally, interaction time per data point differed significantly across conditions, F(2, 61) = 8.60, p < .001, η^2 = 0.22. Participants in the system-randomized condition spent the most time analyzing each data point (M = 114.09, SD = 79.42), compared to the fixed condition (M = 64.71, SD = 35.33) and the user-controlled condition (M = 51.79, SD = 21.53). Fig. 2 illustrates group differences in exploratory behavior.

4.2. Mental model

To test the effect of interface type on participants' objective understanding of the data model, a composite score was calculated based on participants' performance across the three dimensions of objective understanding (feature dependencies, decision prediction, and alternative prediction). The analysis revealed no significant effect of interface type on participants' objective understanding, $F(2, 61) = 0.23$, $p > .05$. Separate analyses of the three dimensions of objective understanding also yielded no significant results (feature dependencies: $F(2, 61) = 1.41$, $p > .05$; decision prediction: $F(2, 61) = 0.37$, $p > .05$; alternative prediction: $F(2, 61) = 0.18$, $p > .05$; see Fig. 3). Overall, average performance on the objective understanding measure was high, with scores showing little variability. Based on the composite score across all understanding dimensions, the mean performance across the three conditions ranged between seven and eight out of possible ten points. There were also no significant differences in confidence ratings between the experimental conditions, $F(2, 61) = 0.64$, $p > .05$. Across all three groups, participants reported being slightly more confident than average in their ability to correctly identify the features that had a relevant influence on the model's prediction. Descriptively, confidence was highest in the fixed condition ($M = 4.55$, $SD = 1.95$). For self-reported understanding a significant effect of interface type was found, $F(2, 61) = 3.33$, $p < .05$, $\eta^2 = 0.10$. It was highest in the fixed condition ($M = 4.27$, $SD = 1.78$) and significantly differed from the user-controlled condition ($M = 3.05$, $SD = 1.47$).

4.3. Satisfaction

Satisfaction with the counterfactual explanations did not differ significantly between the three conditions, according to a one-way ANOVA, $F(2, 61) = 0.40$, $p > .05$. Overall, satisfaction was relatively high across all conditions (for exact values, see Table 1; for group differences see Fig. 4).

4.4. Cognitive workload

We conducted two separate one-way ANOVAs to examine the effect of interface type on both frustration and mental demand, as measured by the corresponding NASA-TLX subscales. Participants reported significantly different levels of frustration across the three conditions, $F(2, 61) = 15.25$, $p < .01$, $\eta^2 = 0.33$. In the user-controlled condition, participants experienced the highest frustration ($M = 12.76$, $SD = 3.03$), which was significantly greater than in the fixed condition ($M = 9.27$, $SD = 3.04$) and system-randomized condition ($M = 7.90$, $SD = 2.74$; see Fig. 4). Likewise, interface type produced significantly different levels of mental demand, $F(2, 61) = 10.89$, $p < .01$, $\eta^2 = 0.26$. Mean mental demand was significantly lower in the fixed condition ($M = 10.09$, $SD = 2.24$) than in the system-randomized condition ($M = 12.42$, $SD = 3.60$) and the user-controlled condition ($M = 14.62$, $SD = 3.55$).

5. Discussion

In this study, we investigated how different modes of parameter control in a counterfactual explanation interface affected human–XAI interaction. Across all conditions, participants could iteratively explore the same set of data points by repeatedly re-generating counterfactual sets; conditions differed only in the locus of control over mutate actions, namely feature weighting and feature exclusion. Specifically, generation parameters were fixed in the fixed condition, assigned automatically in the system-randomized condition, and controlled via sliders in the user-controlled condition. We interpret the findings against the backdrop of prior empirical evidence and discuss priorities for future investigation.

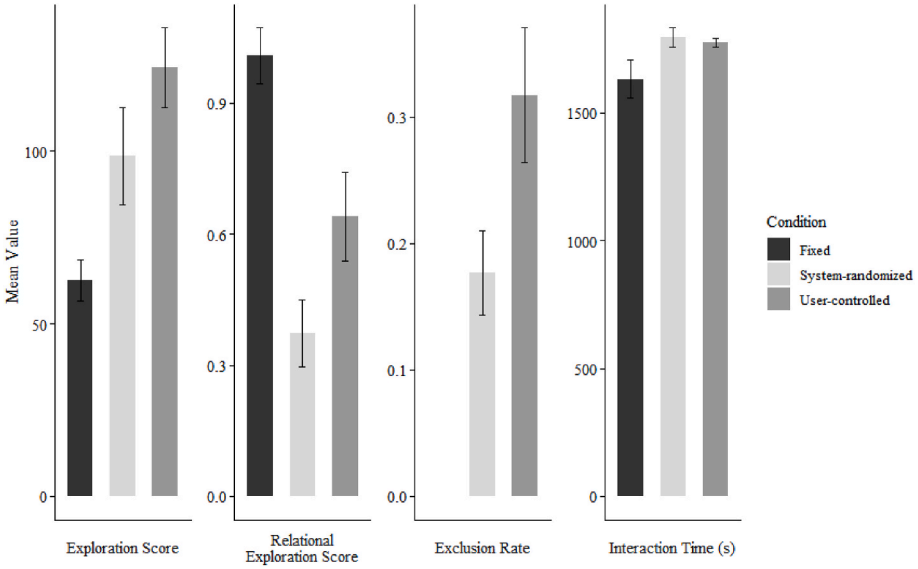


Fig. 2. Mean exploratory behavior measures as a function of experimental condition (fixed, system-randomized, user-controlled). Error bars represent ±1 standard error of the mean. Exclusion rate was not applicable in the fixed condition.

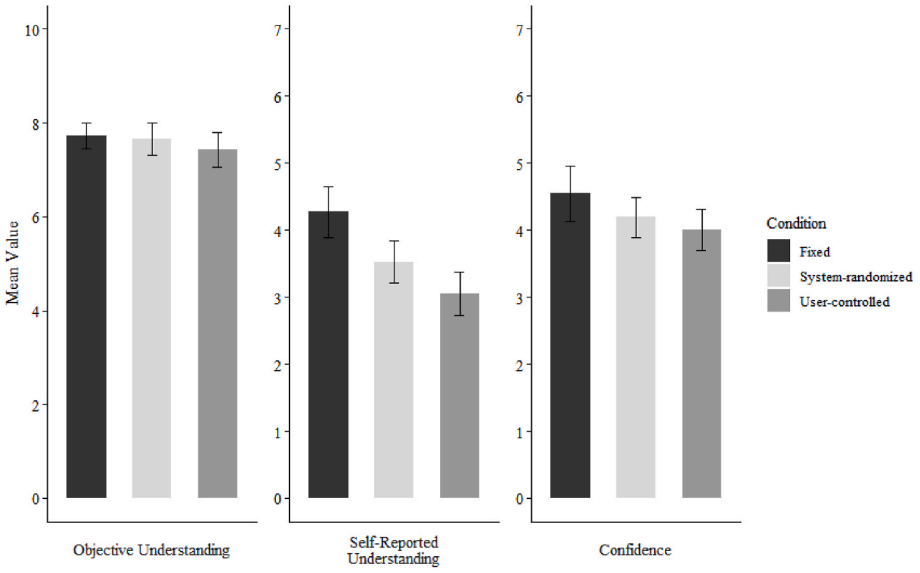


Fig. 3. Participants' objective understanding, self-reported understanding and confidence in their understanding by conditions. Error bars represent ±1 standard error of the mean.

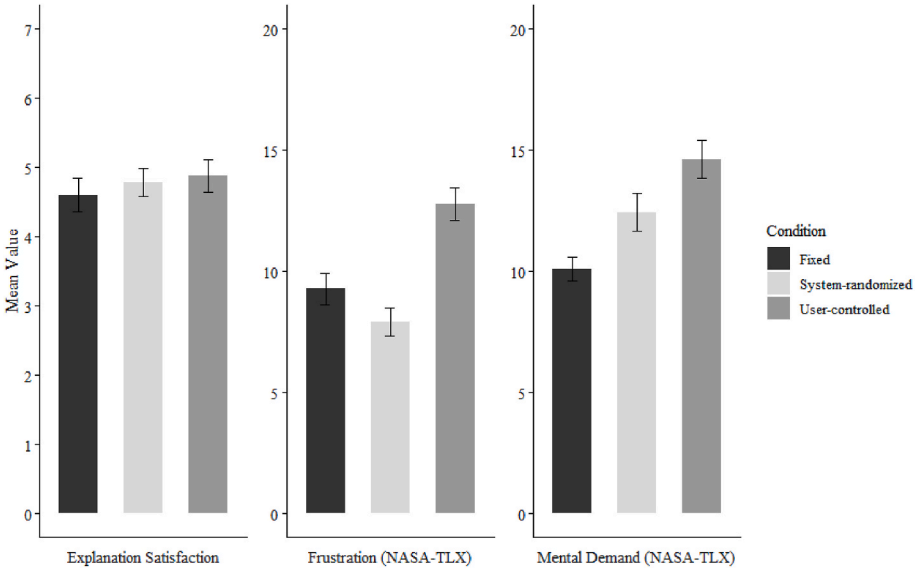


Fig. 4. Participants' explanations satisfaction, frustration and mental demand by conditions. Error bars represent ±1 standard error of the mean.

5.1. User-controlled parameterization increases exploratory breadth, but at the cost of higher subjective cognitive workload

Across conditions, we observed systematic differences in exploratory behavior. Overall exploratory activity, operationalized via an exploration score combining data-point selection and counterfactual generation, was highest in the user-controlled condition. In addition, when participants had direct control over mutate actions, exclusion rates were higher than in the system-randomized condition. We interpret this pattern as a complexity-management strategy. Excluding features reduces the space of potential explanations that users must consider when forming and testing hypotheses about feature–outcome relationships. This interpretation aligns with our workload findings. Participants reported the highest frustration and mental demand in the user-controlled condition, suggesting that active parameter control can impose a substantial operational burden. Prior work similarly suggests that interactive counterfactual interfaces can become demanding when users must coordinate multiple parameter adjustments while trying to attribute changes in counterfactual sets to specific actions (Panigrahi et al., 2025). The elevated frustration documented here likely reflects the dual need to allocate resources to learning how the parameter controls operate while simultaneously analyzing the resulting counterfactuals to infer properties of the underlying model, motivating users to reduce complexity through exclusion.

5.2. System-randomized parameterization promotes in-depth exploration and longer engagement per data point

In-depth exploration, operationalized via the relational exploration score as the ratio of selected data points to generated counterfactual sets, was highest in the system-randomized condition, indicating that participants generated more counterfactual sets per data point. One interpretation is that participants attempted to infer a rationale behind the system-randomized parameterization by generating successive counterfactual sets and comparing resulting patterns. Consistent with this account, interaction time per data point was also highest in the system-

randomized condition. This suggests sustained engagement with the explanations rather than brief trial-and-error. Prior work has linked richer interaction opportunities to longer interaction times (Buçinca et al., 2020; Cheng et al., 2019; Ross et al., 2021). In the present task, which emphasized model understanding rather than time-critical decisions, time spent per data point can be interpreted as an objective indicator of engagement and processing demands, although it does not necessarily map one-to-one onto subjective workload. Notably, the system-randomized condition increased time on task despite lower reported frustration than the user-controlled condition, indicating that time and subjective workload capture partly distinct aspects of the interaction experience. This interpretation is consistent with arguments that longer engagement often co-varies with processing demands (Bertrand et al., 2023).

5.3. Parameter control did not improve mental-model formation or satisfaction in an under-complex abstract setting

Differences in exploratory behavior did not translate into higher explanation satisfaction or better mental models. Neither system-randomized parameter variation nor user-controlled parameterization improved objective understanding relative to the fixed condition. This aligns with mixed findings in the literature. While some studies report gains from interactive explanations (Cheng et al., 2019), others find no such benefits (Bove et al., 2022; Liu et al., 2021). In our data, explanation satisfaction and objective understanding were high across all conditions, which is notable given the elevated workload in the user-controlled condition. Confidence ratings did not track objective performance; descriptively, confidence peaked in the fixed condition and self-reported model understanding was highest in the fixed condition. We argue that the abstract scenario, with relatively few features compared to many applied decision-support contexts, reduced semantic complexity and may have limited the potential marginal benefit of parameter control for learning the model. The resulting pattern, less agency and more confidence, is consistent with an illusion of explanatory depth, a cognitive bias in which people overestimate their understanding of how complex causal systems work (Rozenblit & Keil, 2002). In the context of AI explanations, Chromik and Butz (2021) showed that laypeople in particular exhibit an illusion of explanatory depth, and that this effect can be exacerbated when less information about the system is presented, which aligns with the present descriptive pattern.

5.4. Limitations and future work

As with any study, several limitations warrant consideration. We examined an abstract scenario using a synthetic dataset. This choice allowed us to analyze feature dependencies without confounding world knowledge, but it also meant that the dependencies encoded in the machine-learning model and the task context were under-complex relative to many real-world settings. As a result, comparability to prior work is limited, and transfer to workplace-oriented AI support systems is constrained. Future studies should implement comparable designs in ecologically valid domains and evaluate fixed, system-randomized, and user-controlled parameterization in AI-supported decision making. Such work could also incorporate additional objective, behavior-proximal measures to quantify benefits attributable to parameter control and to distinguish beneficial engagement from unproductive effort.

A second limitation is the exclusively quantitative design. While combining subjective and objective indicators has become standard in human–AI interaction (Silva et al., 2024), mixed-methods approaches remain scarce. Designs that integrate standardized quantitative measures with qualitative process tracing (e.g., think-aloud, interview probes) can yield richer insights into exploration strategies and mental-model formation, thereby supporting a more holistic interpretation and contextualization of results. In the present study, integrating qualitative measures could have provided deeper insight into exploration strategies and might have allowed us to refine analyses by distinguishing relevant subtypes of exploration.

More broadly, our findings show that exploration behavior varies as a function of the locus of control over generation parameters. Although anticipated benefits in understanding and satisfaction did not materialize, we argue for systematically comparing fixed, system-randomized, and user-controlled parameterization in future work while maintaining informational parity. Without carefully matched interfaces and clearly defined control mechanisms, it becomes difficult to attribute observed effects to parameter control per se rather than to general differences in interface richness. Because the effectiveness of parameter control is likely influenced by contextual factors such as domain, task complexity, and user expectations, principled interaction design remains essential (Aljuneidi et al., 2025; Bertrand et al., 2023).

Generalizability is also constrained by the specific counterfactual paradigm and interaction modality studied here. Our interface operationalized parameter control over a DiCE-based generator in a visual dashboard. Other counterfactual mechanisms may expose different control levers and may respond differently to user agency, for example when counterfactuals are constrained for plausibility (Kenny et al., 2021; Smyth & Keane, 2022) and actionability (Poyiadzi et al., 2020), generated under alternative optimization objectives (Dandl et al., 2020), or produced by different model families (Lucic et al., 2022). Moreover, interactivity can be realized through different modalities, including dialogue-based explanation interfaces that support iterative questioning and clarification rather than explicit parameter manipulation (Akula et al., 2022; Sokol & Flach, 2018). Such interfaces may reduce operational burden and may scaffold hypothesis testing differently than slider-based controls. Finally, our study focused on a single-model prediction context. In aggregated-intelligence settings, decisions may be produced by multiple interacting models or agents, which can increase causal complexity and make user understanding more challenging. In these scenarios, interactive visual models that support decomposition of contributions, comparison across agents, and progressive disclosure may become particularly valuable. Future work should therefore replicate our design across counterfactual paradigms, interaction modalities, and multi-agent decision contexts to identify when and for whom user agency yields benefits beyond informational parity.

5.5. Conclusion

This study examined whether different modes of parameter control in counterfactual explanation interfaces, namely fixed parameters, system-randomized parameters, and user-controlled parameters, yield measurable benefits beyond a fixed baseline under informational parity. User-controlled parameterization reliably amplified exploratory breadth and was associated with higher exclusion rates, suggesting that users leveraged control to manage complexity. At the same time, user-controlled parameterization coincided with higher subjective workload and frustration. System-randomized parameterization promoted in-depth exploration and longer engagement per data point, but neither system-randomized nor user-controlled parameterization produced gains in objective understanding or explanation satisfaction relative to the fixed condition. Together, the results suggest that more control is not a sufficient condition for more insight.

A plausible interpretation is that, absent additional informational scaffolding, parameter control primarily changes how much users explore rather than what they learn from that exploration. Users appear to leverage control to simplify the task rather than to refine mental models of the underlying classifier. This aligns with the view that interaction design must manage limited cognitive resources. If an interface increases parameter complexity without commensurate guidance, users' effort may be spent on operating the tool instead of extracting model-relevant regularities.

A practical implication is to treat parameter control as a means to deliver targeted, low-friction insight rather than an end in itself.

Explanation interfaces can be more effective when controls are guided and purposeful, supporting users' inferences about the model rather than merely expanding the space of possible manipulations. In practice, this can involve integrating brief micro-explanations that make the epistemic value of actions transparent (e.g., adjusting weight X helps test hypothesis Y), using progressive disclosure with sensible defaults to bound parameter complexity, and offering example-driven what-if prompts that steer attention toward informative regions of the feature space. Such design choices may preserve the motivational benefits of agency while helping to keep workload in check.

CRediT authorship contribution statement

Lena Kölmel: Writing – review & editing, Writing – original draft, Visualization, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Maximilian Becker:** Writing – review & editing, Software, Methodology, Conceptualization. **Finn Schwall:** Writing – review & editing, Software, Data curation, Conceptualization. **Jutta Hild:** Writing – review & editing, Project administration, Funding acquisition, Conceptualization. **Pascal Birnstill:** Writing – review & editing, Supervision. **Barbara Deml:** Writing – review & editing, Supervision, Resources. **Jürgen Beyerer:** Supervision.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used OpenAI's ChatGPT (Model GPT-4o and GPT-5) and DeepL to improve the language and readability of the manuscript. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

Funding

The study was conducted within the Competence Center KARL–Artificial Intelligence for Work and Learning in the Karlsruhe Region. This research and development project is funded by the Federal Ministry of Research, Technology and Space (BMFTR; 02L19C250) and administered by the Project Management Agency Karlsruhe (PTKA). The authors are solely responsible for the content of this publication.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We thank Franziska Kaste and Blanca Vitallowitz for their assistance with data collection.

Data availability

The dataset used in this study is available at OSF: https://osf.io/jmydt/overview?view_only=7d2cb23f2b8645239b4bc1fc70bec922.

References

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access, 6*, 52138–52160. https://doi.org/10.1109/access.2018.2870052

Akula, A. R., Wang, K., Liu, C., Saba-Sadiya, S., Lu, H., Todorovic, S., Chai, J., & Zhu, S.-C. (2022). CX-ToM: Counterfactual explanations with theory-of-mind for enhancing human trust in image recognition models. *iScience, 25*(1), Article 103581.

Aljuneidi, S., Heuten, W., Wolters, M., & Boll, S. (2025). From explaining to engaging: The effect of interactive AI explanations on citizens' fairness and adoption perceptions. In *Proceedings of the INTERACT 2025: 20th IFIP TC13 international conference on human-computer interaction* (pp. 87–108). Springer. https://doi.org/10.1007/978-3-032-05002-1_5.

Arya, V., Bellamy, R. K. E., Chen, P.-Y., Dhurandhar, A., Hind, M., Hoffman, S. C., Houde, S., Liao, Q. V., Luss, R., Mojsilović, A., Mourad, S., Pedemonte, P., Raghavendra, R., Richards, J., Sattigeri, P., Shanmugam, K., Singh, M., Varshney, K. R., Wei, D., & Zhang, Y. (2019). One explanation does not fit all: A toolkit and taxonomy of AI explainability techniques. *arXiv Preprint arXiv: 1909.03012*. https://arxiv.org/abs/1909.03012.

Becker, M., Vishwesh, V., Birnstill, P., Schwall, F., Wu, S., & Beyerer, J. (2023). RIXA – Explaining artificial intelligence in natural language. In *Proceedings of the 2023 IEEE international conference on data mining workshops (ICDMW)* (pp. 875–884). IEEE. https://doi.org/10.1109/icdmw60847.2023.00118.

Bertrand, A., Viard, T., Belloum, R., Eagan, J. R., & Maxwell, W. (2023). On selective, mutable and dialogic XAI: A review of what users say about different types of interactive explanations. In *Proceedings of the 2023 CHI conference on human factors in computing systems* (pp. 1–21). ACM. https://doi.org/10.1145/3544548.3581314.

Bove, C., Aigrain, J., Lesot, M.-J., Tijus, C., & Detyniecki, M. (2022). Contextualization and exploration of local feature importance explanations to improve understanding and satisfaction of non-expert users. In *In proceedings of the 27th international conference on intelligent user interfaces* (pp. 132–142). ACM. https://doi.org/10.1145/3490099.3511139.

Bove, C., Lesot, M.-J., Tijus, C. A., & Detyniecki, M. (2023). Investigating the intelligibility of plural counterfactual examples for non-expert users: An explanation user interface proposition and user study. In *In proceedings of the 28th international conference on intelligent user interfaces* (pp. 330–340). ACM. https://doi.org/10.1145/3581641.3584082.

Buçinca, Z., Lin, P., Gajos, K. Z., & Glassman, E. L. (2020). Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. In *In proceedings of the 25th international conference on intelligent user interfaces* (pp. 454–464). ACM. https://doi.org/10.1145/3377325.3377498.

Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (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), 1–21. https://doi.org/10.1145/3449287

Byrne, R. M. J. (2019). Counterfactuals in explainable artificial intelligence (XAI): Evidence from human reasoning. In *In proceedings of the twenty-eighth international joint conference on artificial intelligence (IJCAI-19)* (pp. 6276–6282). https://doi.org/10.24963/ijcai.2019/872

Cabrera, Á. A., Tulio Ribeiro, M., Lee, B., Deline, R., Perer, A., & Drucker, S. M. (2023). What did my AI learn? How data scientists make sense of model behavior. *ACM Transactions on Computer-Human Interaction, 30*(1), 1–27. https://doi.org/10.1145/3542921

Cappuccio, E., Esposito, A., Greco, F., Desolda, G., Lanzilotti, R., & Rinzivillo, S. (2025). Explanation user interfaces: A systematic literature review. *arXiv preprint arXiv: 2505.20085*. https://arxiv.org/abs/2505.20085.

Celar, L., & Byrne, R. M. J. (2023). How people reason with counterfactual and causal explanations for artificial intelligence decisions in familiar and unfamiliar domains. *Memory & Cognition, 51*(7), 1481–1496. https://doi.org/10.3758/s13421-023-01407-5

Cheng, F., Liu, D., Du, F., Lin, Y., Zytek, A., Li, H., Qu, H., & Veeramachaneni, K. (2022). VBridge: Connecting the dots between features and data to explain healthcare models. *IEEE Transactions on Visualization and Computer Graphics, 28*(1), 378–388. https://doi.org/10.1109/TVCG.2021.3114836

Cheng, H.-F., Wang, R., Zhang, Z., O'Connell, F., Gray, T., Harper, F. M., & Zhu, H. (2019). Explaining decision-making algorithms through UI. In *In proceedings of the 2019 CHI conference on human factors in computing systems* (pp. 1–12). ACM. https://doi.org/10.1145/3290605.3300789.

Cheng, F., Zouhar, V., Chan, R. S. M., Fürst, D., Strobelt, H., & El-Assady, M. (2024). Interactive analysis of LLMs using meaningful counterfactuals. *arXiv preprint arXiv: 2405.00708*. https://doi.org/10.48550/arXiv.2405.00708

Chromik, M., & Butz, A. (2021). Human-XAI interaction: A review and design principles for explanation user interfaces. In C. Stephanidis, & M. Antona (Eds.), *HCI international 2021 – Late breaking papers: Multimodality, eXtended reality, and artificial intelligence* (pp. 619–640). Cham: Springer. https://doi.org/10.1007/978-3-030-85616-8_36.

Dandl, S., Molnar, C., Binder, M., & Bischl, B. (2020). Multi-objective counterfactual explanations. In *International conference on parallel problem solving from nature* (pp. 448–469). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-58112-1_31.

Das Antar, A., Molaei, S., Chen, Y.-Y., Lee, M. L., & Banovic, N. (2024). VIME: Visual interactive model explorer for identifying capabilities and limitations of machine learning models for sequential decision-making. In *In proceedings of the 37th annual ACM symposium on user interface software and technology (UIST '24)* (pp. 1–21). ACM. https://doi.org/10.1145/3654777.3676323.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2018). Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science, 64*(3), 1155–1170. https://doi.org/10.1287/mnsc.2016.2643

Dix, A., & Ellis, G. (1998). Starting simple: Adding value to static visualisation through simple interaction. In *In proceedings of the working conference on advanced visual interfaces (AVI '98)* (pp. 124–134). ACM. https://doi.org/10.1145/948496.948514.

Dwivedi, R., Dave, D., Naik, H., Singhal, S., Omer, R., Patel, P., Qian, B., Wen, Z., Shah, T., Morgan, G., & Ranjan, R. (2023). Explainable AI (XAI): Core ideas,

techniques, and solutions. *ACM Computing Surveys, 55*(9), 1–33. https://doi.org/10.1145/3561048

Esposito, A., Calvano, M., Curci, A., Greco, F., Lanzilotti, R., & Piccinno, A. (2025). Explanation-driven interventions for artificial intelligence model customization. In C. Santoro, A. Schmidt, M. Matera, & A. Bellucci (Eds.), *End-user development: 10th international symposium, IS-EUD 2025, munich, Germany, June 16–18, 2025, proceedings (lecture notes in computer science, 15713* pp. 161–170). Springer. https://doi.org/10.1007/978-3-031-95452-8_10.

European Union. (2016). Regulation (EU) 2016/679 of the european parliament and of the council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (general data protection regulation). *Official Journal of the European Union, L119*, 1–88. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32016R0679.

Evans, C., & Gibbons, N. J. (2007). The interactivity effect in multimedia learning. *Computers & Education, 49*(4), 1147–1160. https://doi.org/10.1016/j.compedu.2006.01.008

Fails, J. A., & Olsen, D. R. (2003). Interactive machine learning. In *In proceedings of the 8th international conference on intelligent user interfaces (IUI '03)* (pp. 39–45). ACM. https://doi.org/10.1145/604045.604056.

Foley, J., van Dam, A., Feiner, S., & Hughes, J. (Eds.). (1996). *Computer graphics: Principles and practice* (2nd ed.). Addison-Wesley Professional. https://doi.org/10.5860/choice.51-2713.

Gómez, O., Holter, S., Yuan, J., & Bertini, E. (2020). ViCE: Visual counterfactual explanations for machine learning models. In *In proceedings of the 25th international conference on intelligent user interfaces (IUI '20)* (pp. 531–535). ACM. https://doi.org/10.1145/3377325.3377536.

Gómez, O., Holter, S., Yuan, J., & Bertini, E. (2021). AdViCE: Aggregated visual counterfactual explanations for machine learning model validation. In *2021 IEEE visualization conference (VIS)* (pp. 31–35). IEEE. https://doi.org/10.1109/VIS49827.2021.9623271.

Gómez-Carmona, O., Casado-Mansilla, D., López-de-Ipiña, D., & García-Zubia, J. (2024). Human-in-the-loop machine learning: Reconceptualizing the role of the user in interactive approaches. *Internet of Things, 25*, Article 101048. https://doi.org/10.1016/j.iot.2023.101048

Guidotti, R. (2024). Counterfactual explanations and how to find them: Literature review and benchmarking. *Data Mining and Knowledge Discovery, 38*(5), 2770–2824. https://doi.org/10.1007/s10618-022-00831-6

Hart, S. G. (2006). NASA-task load index (NASA-TLX); 20 years later. *Proceedings of the Human Factors and Ergonomics Society - Annual Meeting, 50*(9), 904–908. https://doi.org/10.1177/154193120605000909

Hernandez-Bocanegra, D. C., & Ziegler, J. (2021). Conversational review-based explanations for recommender systems: Exploring users' query behavior. In *In proceedings of the 3rd conference on conversational user interfaces (CUI '21)* (pp. 1–11). ACM. https://doi.org/10.1145/3469595.3469596.

Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2023). Measures for explainable AI: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-AI performance. *Frontiers of Computer Science, 5*, Article 1096257. https://doi.org/10.3389/fcomp.2023.1096257

Hohman, F., Head, A., Caruana, R., DeLine, R., & Drucker, S. M. (2019). Gamut: A design probe to understand how data scientists understand machine learning models. In *In proceedings of the 2019 CHI conference on human factors in computing systems (CHI '19)*. https://doi.org/10.1145/3290605.3300809. Paper 579, 1–13). ACM.

Holzinger, A., Plass, M., Kickmeier-Rust, M., Holzinger, K., Crişan, G. C., Pintea, C.-M., & Palade, V. (2019). Interactive machine learning: Experimental evidence for the human in the algorithmic loop. *Applied Intelligence, 49*(7), 2401–2414. https://doi.org/10.1007/s10489-018-1361-5

Hudon, A., Demazure, T., Karran, A., Léger, P.-M., & Sénécal, S. (2021). Explainable artificial intelligence (XAI): How the visualization of AI predictions affects user cognitive load and confidence. In F. D. Davis, R. Riedl, J. vom Brocke, P.-M. Léger, A. B. Randolph, & G. Müller-Putz (Eds.), *Information systems and neuroscience* (pp. 237–246). Springer. https://doi.org/10.1007/978-3-030-88900-5_27.

IBM Corp. (2020). *IBM SPSS statistics for windows (Version Version 27.0)*. IBM Corp [Computer software].

Islam, M. R., Ahmed, M. U., Barua, S., & Begum, S. (2022). A systematic review of explainable artificial intelligence in terms of different application domains and tasks. *Applied Sciences, 12*(3), 1353. https://doi.org/10.3390/app12031353

Janlert, L.-E., & Stolterman, E. (2017). The meaning of interactivity—Some proposals for definitions and measures. *Human-Computer Interaction, 32*(3), 103–138. https://doi.org/10.1080/07370024.2016.1226139

Jeyasothy, A., Laugel, T., Lesot, M.-J., Marsala, C., & Detyniecki, M. (2022). Integrating prior knowledge in post-hoc explanations. In D. Ciucci, I. Couso, J. Medina, D. Ślęzak, D. Petturiti, B. Bouchon-Meunier, & R. R. Yager (Eds.), *Information processing and management of uncertainty in knowledge-based systems (CCIS, 1602* pp. 707–719). Springer. https://doi.org/10.1007/978-3-031-08974-9_56.

Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness*. Harvard University Press.

Kalasampath, K., Spoorthi, K. N., Sajeev, S., Kuppa, S. S., Ajay, K., & Maruthamuthu, A. (2025). A literature review on applications of explainable artificial intelligence (XAI). *IEEE Access, 13*, 41111–41140. https://doi.org/10.1109/ACCESS.2025.3546681

Kenny, E. M., & Keane, M. T. (2021). On generating plausible counterfactual and semi-factual explanations for deep learning. *Proceedings of the AAAI Conference on Artificial Intelligence, 35*(13), 11575–11585. https://doi.org/10.1609/aaai.v35i13.17377

Koh, S., Kim, B. H., & Jo, S. (2025). Understanding the user perception and experience of interactive algorithmic recourse customization. *ACM Transactions on Computer-Human Interaction, 31*(3), 1–25. https://doi.org/10.1145/3674503. Article 43.

Kulesza, T., Burnett, M., Wong, W.-K., & Stumpf, S. (2015). Principles of explanatory debugging to personalize interactive machine learning. In *In proceedings of the 20th international conference on intelligent user interfaces (IUI '15)* (pp. 126–137). ACM. https://doi.org/10.1145/2678025.2701399.

Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, L., Schmidt, E., Sesing, A., & Baum, K. (2021). What do we want from explainable artificial intelligence (XAI)? A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artificial Intelligence, 296*, Article 103473. https://doi.org/10.1016/j.artint.2021.103473

Liao, Q. V., Gruen, D., & Miller, S. (2020). Questioning the AI: Informing design practices for explainable AI user experiences. In *In proceedings of the 2020 CHI conference on human factors in computing systems (CHI '20)* (pp. 1–15). ACM. https://doi.org/10.1145/3313831.3376590.

Liu, H., Lai, V., & Tan, C. (2021). Understanding the effect of out-of-distribution examples and interactive explanations on human–AI decision making. *Proceedings of the ACM on Human-Computer Interaction, 5*(CSCW2), 1–45. https://doi.org/10.1145/3479552

Lucic, A., Oosterhuis, H., Haned, H., & Rijke, M. de (2022). FOCUS: Flexible optimizable counterfactual explanations for tree ensembles. *Proceedings of the AAAI Conference on Artificial Intelligence, 36*(5), 5313–5322. https://doi.org/10.1609/aaai.v36i5.20468

Martijn, M., Conati, C., & Verbert, K. (2022). "Knowing me, knowing you": Personalized explanations for a music recommender system. *User Modeling and User-Adapted Interaction, 32*(1–2), 215–252. https://doi.org/10.1007/s11257-021-09304-9

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence, 267*, 1–38. https://doi.org/10.1016/j.artint.2018.07.007

Miller, T., Howe, P., & Sonenberg, L. (2017). Explainable AI: Beware of inmates running the asylum, or: How I learnt to stop worrying and love the social and behavioural sciences. *arXiv preprint arXiv:1712.00547*. https://arxiv.org/abs/1712.00547.

Mothilal, R. K., Sharma, A., & Tan, C. (2020). Explaining machine learning classifiers through diverse counterfactual explanations. In *In proceedings of the 2020 conference on fairness, accountability, and transparency (FAccT '20)* (pp. 607–617). ACM. https://doi.org/10.1145/3351095.3372850.

Myers, C. M., Freed, E., Laris Pardo, F., & Furqan, A. (2020). Revealing neural network bias to non-experts through interactive counterfactual examples. arXiv preprint arXiv:2001.02271 https://arxiv.org/abs/2001.02271.

Naveed, S., Stevens, G., & Robin-Kern, D. (2024). An overview of the empirical evaluation of explainable AI (XAI): A comprehensive guideline for user-centered evaluation in XAI. *Applied Sciences, 14*(23), Article 11288. https://doi.org/10.3390/app142311288

Panigrahi, I., Kim, S. S. Y., Liaqat, A., Jinturkar, R., Russakovsky, O., Fong, R., & Abtahi, P. (2025). Interactivity × explainability: Toward understanding how interactivity can improve computer vision explanations. In *Extended abstracts of the 2025 CHI conference on human factors in computing systems (CHI EA '25)*ACM. https://doi.org/10.1145/3706599.3719730, 1–9.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830. http://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf.

Poyiadzi, R., Sokol, K., Santos-Rodriguez, R., De Bie, T., & Flach, P. (2020). FACE: Feasible and actionable counterfactual explanations. In *In Proceedings of the 2020 AAAI/ACM conference on AI, ethics, and society (AIES '20)* (pp. 344–350). ACM. https://doi.org/10.1145/3375627.3375850.

Raees, M., Meijerink, I., Lykourentzou, I., Khan, V.-J., & Papangelis, K. (2024). From explainable to interactive AI: A literature review on current trends in human–AI interaction. *International Journal of Human-Computer Studies, 189*, Article 103301. https://doi.org/10.1016/j.ijhcs.2024.103301

Ross, A., Chen, N., Hang, E. Z., DeLine, R., Glassman, E. L., & Doshi-Velez, F. (2021). Evaluating the interpretability of generative models by interactive reconstruction. In *In proceedings of the 2021 CHI conference on human factors in computing systems (CHI '21)* (pp. 1–15). ACM. https://doi.org/10.1145/3411764.3445296.

Roussou, M. (2004). Learning by doing and learning through play: **An** exploration of interactivity in virtual environments for children. *Computer Entertainment, 2*(1), 10–33. https://doi.org/10.1145/973801.973818

Rozenblit, L., & Keil, F. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science, 26*(5), 521–562. https://doi.org/10.1207/s15516709cog2605_1

Shang, R., Feng, K. J. K., & Shah, C. (2022). Why am I not seeing it? Understanding users' needs for counterfactual explanations in everyday recommendations. In *In proceedings of the 2022 ACM conference on fairness, accountability, and transparency (FAccT '22)* (pp. 1330–1340). ACM. https://doi.org/10.1145/3531146.3533189.

Silva, A., Schrum, M., Hedlund-Botti, E., Gopalan, N., & Gombolay, M. (2023). Explainable artificial intelligence: Evaluating the objective and subjective impacts of XAI on human–agent interaction. *International Journal of Human-Computer Interaction, 39*(7), 1390–1404. https://doi.org/10.1080/10447318.2022.2101698

Silva, A., Tambwekar, P., Schrum, M., & Gombolay, M. (2024). Towards balancing preference and performance through adaptive personalized explainability. In *In proceedings of the 2024 ACM/IEEE international conference on human-robot interaction (HRI '24)* (pp. 658–668). https://doi.org/10.1145/3610977.3635000. ACM.

Skulmowski, A. (2024). Learning by doing or doing without learning? The potentials and challenges of activity-based learning. *Educational Psychology Review, 36*(1), 1–26. https://doi.org/10.1007/s10648-024-09869-y

Smyth, B., & Keane, M. T. (2022). A few good counterfactuals: Generating interpretable, plausible and diverse counterfactual explanations. In *International conference on case-based reasoning* (pp. 18–32). Cham: Springer International Publishing.

Sokol, K., & Flach, P. A. (2018). Glass-Box: Explaining AI decisions with counterfactual statements through conversation with a voice-enabled virtual assistant. *IJCAI, 18*, 5868–5870.

Speckmann, P., Nadj, M., & Janiesch, C. (2025). Interactive explainable intelligent systems: Requirements, design principles, and prototypical implementation. In S. Chatterjee, J. vom Brocke, & R. Anderson (Eds.), *Local solutions for global challenges: 20th international conference on design science research in information systems and technology, DESRIST 2025, Montego Bay, Jamaica, June 2–4, 2025, proceedings, part II (Lecture notes in computer science, 15704* pp. 51–65). Springer. https://doi.org/10.1007/978-3-031-93979-2_4.

Suffian, M., Kuhl, U., Bogliolo, A., & Alonso-Moral, J. M. (2025). The role of user feedback in enhancing understanding and trust in counterfactual explanations for explainable AI. *International Journal of Human-Computer Studies, 199*, Article 103484. https://doi.org/10.1016/j.ijhcs.2025.103484

Tullio, J., Dey, A. K., Chalecki, J., & Fogarty, J. (2007). How it works: A field study of non-technical users interacting with an intelligent system. In *Proceedings of the SIGCHI conference on human factors in computing systems (CHI '07)* (pp. 31–40). ACM. https://doi.org/10.1145/1240624.1240630.

Turchi, T., Malizia, A., Paternò, F., Borsci, S., & Chamberlain, A. (2024). Adaptive XAI: Towards intelligent interfaces for tailored AI explanations. In *In companion proceedings of the 29th international conference on intelligent user interfaces (IUI '24 companion)* (pp. 119–121). ACM. https://doi.org/10.1145/3640544.3645253.

Ullah, N., Khan, J. A., De Falco, I., & Sannino, G. (2025). Explainable artificial intelligence: Importance, use domains, stages, output shapes, and challenges. *ACM Computing Surveys, 57*(4), 1–36. https://doi.org/10.1145/3705724

Verma, S., Boonsanong, V., Hoang, M., Hines, K., Dickerson, J., & Shah, C. (2024). Counterfactual explanations and algorithmic recourses for machine learning: A review. *ACM Computing Surveys, 56*(12), 1–42. https://doi.org/10.1145/3677119

Wang, X., & Yin, M. (2021). Are explanations helpful? A comparative study of the effects of explanations in AI-assisted decision-making. In *In proceedings of the 26th international conference on intelligent user interfaces (IUI '21)* (pp. 318–328). ACM.

Warren, G., Smyth, B., & Keane, M. T. (2022). "Better" counterfactuals, ones people can understand: Psychologically-plausible case-based counterfactuals using categorical features for explainable AI (XAI). In M. T. Keane, & N. Wiratunga (Eds.), *Case-based reasoning research and development: 30th international conference, ICCBR 2022, Nancy, France, September 12–15, 2022, proceedings (Lecture notes in computer science, 13405* pp. 63–78). Springer. https://doi.org/10.1007/978-3-031-14923-8_5.

Weld, D. S., & Bansal, G. (2019). The challenge of crafting intelligible intelligence. *Communications of the ACM, 62*(6), 70–79. https://doi.org/10.1145/3282486

Xie, Y., Chen, M., Kao, D., Gao, G., & Chen, X. A. (2020). CheXplain: Enabling physicians to explore and understand data-driven, AI-enabled medical imaging analysis. In *In proceedings of the 2020 CHI conference on human factors in computing systems (CHI '20)* (pp. 1–13). https://doi.org/10.1145/3313831.3376807. ACM.

Yi, J. S., Kang, Y. A., Stasko, J., & Jacko, J. (2007). Toward a deeper understanding of the role of interaction in information visualization. *IEEE Transactions on Visualization and Computer Graphics, 13*(6), 1224–1231. https://doi.org/10.1109/TVCG.2007.70515

Zhang, Y., Liao, Q. V., & Bellamy, R. K. E. (2020). Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *In proceedings of the 2020 conference on fairness, accountability, and transparency (FAccT '20)* (pp. 295–305). ACM. https://doi.org/10.1145/3351095.3372852.