On Selective, Mutable and Dialogic XAI: a Review of What Users Say about Different Types of Interactive Explanations

Astrid Bertrand, Tiphaine Viard, Rafik Belloum, James R Eagan, Winston Maxwell

On Selective, Mutable and Dialogic XAI: a Review of What Users Say about Different Types of Interactive Explanations

Astrid Bertrand
LTCI, Télécom Paris, Institut Polytechnique de Paris
Palaiseau, France
astrid.bertrand@telecom-paris.fr

Tiphaine Viard
i3, Télécom Paris, Institut Polytechnique de Paris
Palaiseau, France
tiphaine.viard@telecom-paris.fr

Rafik Belloum
Univ. Polytechnique Hauts-de-France
LAMIH, CNRS, UMR 8201
F-59313 Valenciennes, France
& LTCI, Télécom Paris, Institut Polytechnique de Paris
Palaiseau, France
rafik.belloum@uphf.fr

James R. Eagan
LTCI, Télécom Paris, Institut Polytechnique de Paris
Palaiseau, France
james.eagan@telecom-paris.fr

Winston Maxwell
i3, Télécom Paris, Institut Polytechnique de Paris
Palaiseau, France
winston.maxwell@telecom-paris.fr

ABSTRACT

Explainability (XAI) has matured in recent years to provide more human-centered explanations of AI-based decision systems. While static explanations remain predominant, interactive XAI has gathered momentum to support the human cognitive process of explaining. However, the evidence regarding the benefits of interactive explanations is unclear. In this paper, we map existing findings by conducting a detailed scoping review of 48 empirical studies in which interactive explanations are evaluated with human users. We also create a classification of interactive techniques specific to XAI and group the resulting categories according to their role in the cognitive process of explanation: "selective", "mutable" or "dialogic". We identify the effects of interactivity on several user-based metrics. We find that interactive explanations improve perceived usefulness and performance of the human+AI team but take longer. We highlight conflicting results regarding cognitive load and overconfidence. Lastly, we describe underexplored areas including measuring curiosity or learning or perturbing outcomes.

CCS CONCEPTS

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

KEYWORDS

interactivity, explainability, interpretability, human-grounded evaluations, artificial intelligence

CHI '23, April 23–28, 2023, Hamburg, Germany

ACM ISBN 978-1-4503-9421-5/23/04…$15.00
https://doi.org/10.1145/3544548.3581314

ACM Reference Format:
Astrid Bertrand, Tiphaine Viard, Rafik Belloum, James R. Eagan, and Winston Maxwell. 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 (CHI '23), April 23–28, 2023, Hamburg, Germany.* ACM, New York, NY, USA, 22 pages. https://doi.org/10.1145/3544548.3581314

1 INTRODUCTION

Explainability (XAI) — the practice of explaining the inner workings of artificial intelligence systems — is a major challenge from legal, social and technical standpoints. XAI has seen a surge of interest in recent years, with multiple contributions across fields [1]. Although many explainability techniques have been designed, questions remain about how explanations can be best communicated to users. Notably, recent work focusing on human-centered explanations has been advocating the design of interactive explanations [2, 9, 67, 77, 122], which is considered more effective based in large part on how people explain things to each other [77]. For example, people expect explanations to be provided in a personalized request-response pattern [27]. According to Hesslow [45] and Lipton [72], causal explanations are usually presented in relation to a specific "foil", *i.e.* a contrast. One does not ask "why P?" but rather "why P *and not Q*?". To provide meaningful explanations, the explanation agent has then to find, for each user, the adequate foil. Interactivity is one way to learn what that foil is, by iteratively collecting user information. Research in the field of education shows that interactivity plays a fundamental role in learning [12, 105]. Barker *et al* describe interactivity as "a necessary and fundamental mechanism for knowledge acquisition" [12].

However, the term "interactive" has multiple meanings in the XAI community, referring to different kinds of user interactions. According to Miller [77], the ideal interaction model follows a human-like dialogue structure, where the AI agent is able to answer a series of questions. Other types of user interaction have been implemented by XAI researchers, such as simulating the black box with new inputs [22, 23, 81], re-configuring the explanation

space [47], changing explanations [55, 110], etc. However, these studies in XAI do not use a common vocabulary to designate different interaction types, making it difficult to study and draw general conclusions on the different forms of interactive XAI.

The visualization (Infovis) [5, 53, 95, 125, 128] and other Human-Computer Interaction (HCI) [92, 105] communities have done extensive work on the classification of different modes of interaction. The XAI field is less mature. We believe that the XAI field would benefit from using a more precise and shared vocabulary to designate the different types of interactivity, taking inspiration from other HCI sub-fields.

In this article, we conduct a detailed scoping review on interactive and user-evaluated explainability systems. We survey two popular digital libraries for the HCI community: IEEE Xplore and ACM Digital Library. Our contributions are the following:

- We adapt existing HCI taxonomies of interactivity types to XAI,
- We analyze how interaction techniques support the human cognitive process of explaining,
- We analyze the extent, nature and distribution of the interactive XAI systems included in the review,
- We offer a summary of the user-based evaluation metrics implemented in interactive XAI,
- We offer a qualitative summary of the effects of interactive explanations on several user-based evaluation metrics.

Our work is guided by 4 research questions.

RQ1: What are the interactivity approaches that have been implemented so far in the XAI field? First, we want to identify the types of interactivity that have been explored, drawing on both HCI and XAI literature, and that have been evaluated by users. This allows us to see if existing taxonomies from the HCI and Infovis domains can be used and adapted to the XAI domain. We then propose a two-level taxonomy of interaction techniques in XAI.

RQ2: In what context, with what content, and in what form were the interactive explanations presented to users? Second, we want to understand in what contexts and with what methods interactive explanations have been implemented in the literature. Due to the increasingly large number of articles on XAI, researchers may be overlooking best practices and opportunities for interaction. Interactive explanations can be presented to users in multiple ways: they can include on demand information, ask the user for feedback, take different forms and contain vastly different amounts of information. To illustrate the complexity of designing interactions, Sims [105] referred to it as "an art" requiring multiple considerations and a vast array of skills on the part of designers. This article aims at helping XAI system builders by centralizing examples of interactive explanations taken from various contexts (user expertise, XAI method, domain...). To that end, we examine the characteristics of interactive explanations that have been implemented and evaluated with users. We provide a qualitative analysis of the context (domain, audience, data type), content (explanation focus and method) and communication types (interactivity and representation) found in our scoping review.

RQ3: What are the metrics used in user-based evaluations of interactive explanations? Third, we want to report how XAI researchers have been measuring their explanation systems based on human-grounded evaluations [32]. To the best of our knowledge, there have been few efforts to list user-based evaluation measures for explainability [46], and this work is the first in interactive XAI.

RQ4: What are the effects of interactive explanations on users' perception of explanations? Finally, we want to identify the different effects interactive explanations have on the user experience. Over the past few years, a growing body of work has been testing interactive XAI systems with real users, generating sometimes contradictory findings. Cheng et al. [22] find that the possibility to simulate new predictions by changing input features improved user understanding compared to static explanations. However, concerns were expressed in [73] because interactive explanations were found to reinforce users' over reliance on AI suggestions. One possibility is that interactive explanations were more complex to interpret in [73]'s study, leading to information overload. Similarly, [16] and [42] found no statistically significant improvement of interactive explanations over static explanations with respect to comprehensibility and satisfaction, respectively. We present a qualitative summary of the effects of interactive explanations on XAI's many goals, *i.e.* user trust, user satisfaction, understanding of the AI model, performance at a task, perceived fairness etc. This work is the first, to our knowledge, to summarize the effects of interactive XAI from a user perspective through a scoping survey, paving the way for following systematic reviews to formally disentangle these results.

The remainder of the paper is organized as follows: we start by describing the relevant related work in Section 2, before laying out our methodology in Section 3. This leads us to outlining interactivity types for XAI in Section 4, and analyzing the papers we surveyed in Section 5. Finally, Sections 6, 7 and 8 are dedicated, respectively, to discussing open challenges for interactive XAI, highlighting the limits of our work and concluding.

2 BACKGROUND AND RELATED WORKS

Below, we highlight work in HCI, XAI, and education that is relevant to this paper. We also highlight, through these different strands of literature, reasons to believe that interactivity in XAI could help users in building sense and knowledge about models.

2.1 Interactivity in HCI

Defining interactivity proves challenging, and multiple definitions have been offered over time. Early work on interactivity defined it simply as the extent to which a user can "activate" [105] or "exert an influence" [112, 114] on the technology being used, its form and its content. In 1997, Sims [105] mentioned that "there appears to be no consensus of what interactivity actually represents or involves". Dix et al. [29] and Foley et al. [37] broadly define it using the keywords "communication between user and system" and "human-computer dialogue" [128]. In Infovis, Yi et al. [128] view interaction techniques as "the features that provide users with the ability to directly or indirectly manipulate and interpret representations". The authors noted that Infovis systems were designed to communicate information from the computer to the user, but less so for the user to enter data, thus overlooking an entire aspect of interaction in

HCI. Therefore, differences arise between HCI subdomains on how interactivity is defined. At first glance, it seems that the vision adopted by the Infovis domain could correspond to interactivity in XAI. In the XAI field as well, the user needs to manipulate, interpret and discover information about the model from explanations or raw data. In Section 4, we will examine how adapted the Infovis' view of interaction is to the XAI domain. Despite the lack of a consensual definition, Janlert et al. [49] state that "there seems to be a common sense understanding of interactivity as something fairly simple" that HCI researchers see as "the control and action between a human and an artifact or system."

However, defining the different types of interactions quickly complicates the task. Some studies have addressed it by proposing taxonomies of user-system interactions. Early ones attempted to provide holistic views of the interaction space in HCI; they focused on interaction *levels*, with the idea that "the higher the interaction level, the better the product" [105]. For example, Rhodes and Azbell [92] introduced a three-level scale of interactivity, ranging from *reactive* to *proactive* to *coactive*. Schwier and Misanchuk [101] added two other dimensions to this taxonomy: functions (confirmation, pacing, navigation, inquiry, elaboration) and transactions (keyboard, touch screen, mouse, voice). Sims' taxonomy [105] extends the two previous ones by intertwining functions and levels. It is presented as a scale from basic to complex with the following levels of interactivity: *object, linear, hierarchical, support, update, construct, reflective, simulation, hyperlinked, nonimmersive contextual and immersive virtual*. In the Infovis domain, there is typically no hierarchy between interaction types; however, taxonomies with finer granularity have been designed. For example, Yi et al. [128] observes a difference of approach between system-centric taxonomies (including categories like "interactive linking and brushing" [53] or "navigating", *e.g.* zooming, panning [125]) and user-task-centric taxonomies (including categories like "compare within relations" [95] or "retrieve value" [5]). The taxonomy in [128] proposes to "connect user objectives with the interaction techniques that help accomplish them." It includes seven categories: *select, explore, reconfigure, encode, abstract/elaborate, filter, connect*. Yi et al.'s taxonomy has been extensively used and referred to in Infovis in the last decade.

2.2 Interactivity in Explainability

The call for more interactive explanations in XAI finds roots in results from the social sciences about how people communicate explanations and in the growing number of studies focusing on human needs rather than solely technical aspects.

Works such as Miller's review [77] have been putting to the foreground results from philosophy and cognitive sciences to guide the design of explainable systems. He notably finds that "an explanation is an interaction between two roles: explainer and explainee". As such explanations should be thought as a social process, *i.e.* a conversation. He also mentions the rules that govern this interaction such as Grice's maxims [40] of quality (say only what is true), quantity (say no more than you need to), relation (say what is relevant to the conversation) and manner (say it in a nice way). Although it is easier to imagine these exchanges taking place in natural language, Miller argues that this interaction can use other media such as images, keywords, or logical rules, while still respecting Grice's maxims. This work defines what "human-like" explanations should look like, arguing that users of XAI systems will expect explanations to be delivered in this manner.

The line of research on interactive XAI was also spurred by a call for personalized explanations, following the results highlighted by Miller. Work pertaining to the technical aspects of XAI also identify the importance of such "user-centric" explanations. [100, 108]. Numerous papers have emphasized the need for explanations that are tailored to the context, audience and purpose of the explanation [2, 32, 33, 36, 90]. Scheinder and Handali [100] reviewed XAI studies focusing on personalization. For each paper in their corpus, they documented personalized explanation properties (complexity, content and presentation), personalization granularity (to each user or per category of user) and personalization automation (manual or automatic). Additionally, they observed that personalization of explanations can be either iterative or one-off, with user information being collected once prior to showing explanations [100, 108]. While the personalization of explanations is particularly important given the role of explanations in filling one's specific knowledge gaps, we believe there is a greater granularity of interaction to explore beyond the categories mentioned in [100].

More and more HCI researchers have been investigating user's needs for XAI using standard HCI methods [60, 71, 88, 113]. These efforts have resulted in numerous examples of sophisticated interactive interfaces integrating sometimes complex XAI techniques. For example, the strand of research called "conversational XAI" made significant strides in providing explanations in natural language to a wide range of user questions [43, 44, 108].

2.3 Interactivity for learning and sensemaking

XAI is also deeply connected to results in educational research. The parallel seems natural as the XAI field aspires machines to teach humans [100] or machines to explain themselves. According to Roussou [97], many educational researchers agree that interactivity plays an important role in learning, notably by supporting "learning by doing". Amthor [6] argues that "people retain about 20% of what they hear; 40% of what they see and hear; and 75% of what they see, hear, and do.". This follows the constructivist approach, which emphasizes the need for people to build knowledge by testing and simulating new situations that have meaning for them [28, 97]. Kent *et al.* [54] demonstrates through quantitative user studies "the role of interactivity as a process of knowledge construction" and further asserts that interactivity patterns inform on the actual learning process of an individual. Evans and Gibbons [34] finds that interactivity promotes deep learning by stimulating users' cognitive engagement in the learning process. To tie more concretely these results to the XAI field, we can draw a parallel between the processes of learning, knowledge construction and that, closely related, of sensemaking. Cabrera *et al.* studied the cognitive process of sensemaking of models, and highlighted that "understanding of models is an iterative and ongoing process", motivating the need for their XAI system to be interactive. In this case, the sensemaking—or knowledge construction—, comes from the ability to iterate between the discovery of instances, the formation of hypotheses, their evaluation, etc.

3 SURVEY METHODOLOGY

To review the role of interactivity in XAI, we conducted a scoping review drawn from an initial extraction of 716 papers, narrowed down to our final corpus comprising 48 articles. In this section we detail the characteristics and different phases of the survey method.

3.1 Review type

Like a systematic review [82], a scoping review [8] includes many rigorous steps to survey the literature. Scoping reviews do not require the pre-registration of the results nor the assessment of the quality of the studies [83] as systematic reviews do, but they include similar methodological steps: the definition of research questions, a systematized search and selection process, and an analysis and reporting the results [8]. We followed the standardized search and selection methods from the systematic review methodologies [87], as suggested in [8] for scoping reviews, to ensure the replicability and transparency of our findings. In particular, we followed the following steps of the PRISMA methodology outlined in [87]: paper identification, screening, eligibility evaluation and analysis procedure. This allows us to guarantee the quality of our search and selection process, as encouraged by the PRISMA Extension for Scoping Reviews PRISMA-ScR [118].

Scoping reviews are an appropriate survey type to examine how research is conducted on a specific topic, give a summary of the focus of the field, map key concepts, identify the types of evidence found in a field, pave the way for future systematic reviews, and identify gaps in the literature [83]. This corresponds to the objectives of this paper: identify, map, report and discuss the available evidence on interactivity in XAI.

However, our work goes beyond what is traditionally expected of a scoping review in particular in Section 5.3, where we advance a summary of the effects of interactivity through Figure 3. We argue this step enables us to better delimit gaps in the literature, and provide qualitative grounds for a following systematic review on a more restricted set of studies. This analysis is made possible through a minimal quality control of the included studies that we enforced through the exclusion of entries that were not published in a peer-reviewed conference proceeding or journal. However, a more thorough quality assessment of studies - which entails a restriction on the scope of the survey - should be performed in order to extract quantitative evidence about the effects of interactivity. Here, we aim at identifying the different types of results in the interactive XAI field and orientate further research. Section 7 discusses the limitation of the methodology in further detail.

For all these reasons, we refer to our type of review as a detailed scoping review.

3.2 Corpus creation

3.2.1 Identification. We focused on the ACM Digital Library and IEEE Xplore, two popular databases for the HCI community, which encompass prominent publishing venues for the XAI field (ACM CHI, ACM IUI, IEEE VIS, IEEE TVCG...). Consequently, we focused on XAI work that mainly—though not exclusively—pertain to the HCI community, rather than the computer science side of XAI. The main reason for this is that our focus was on interactivity and user studies—two topics finding roots in HCI. Moreover, the CS side of XAI has been historically and predominantly occupied with technical advances in XAI [32], and has only very recently taken into consideration the user's perspective. While we acknowledge that more interactive XAI systems have been emerging from the CS community recently, such as [106], interaction design has been quite distant from theoretical domains in computer science, as mentioned in [1]. This led us to focus on HCI databases and leave out works published in purely AI conferences, such as NeurIPS, AAAI, or CVPR, among others.

Our aim was to review different types of interactive explanations, focusing on how they are perceived by end users. Therefore, we narrowed our focus to work presenting an XAI interface and including a user-based evaluation of the XAI system. Note that there also exist non user-based evaluations of XAI methods. Finale Doshi-Velez and Been Kim [32] distinguish three evaluation strategies: application-grounded—testing explanations in real-word settings with domain experts—, human-grounded—testing explanations with lay users—, and functionality-grounded—testing explanations using metrics that do not require human feedback. The scope of our survey is limited to empirical studies with human subjects, as we are interested on the users' perception of XAI systems. Providing insight into how people interact with XAI can guide practitioners in making more effective technical and design choices.

The keyword search was contextualized focusing on three dimensions: *AI Systems, Explainability* and *User studies*. The term "interaction" is ubiquitous in HCI (for example the CSS concepts section in ACM papers often include the term), and as such we did not restrict our keyword search to this dimension, choosing instead to select articles on interactive explanations in the eligibility phase. Since we wanted to focus on articles whose main topic was AI, we searched for keywords representing AI systems and explainability dimensions in the Title, Abstract and Author Keywords fields. For the user study dimension, we searched the full text of the articles: we noticed that often, authors do not explicitly mention that they conducted a user-based evaluation in their abstract. The search results were limited to relatively recent articles (2015 or later), as XAI is a recent field of study, found to be expanding around 2016-2017 [2, 13]. In addition, user-based evaluations and interest from the HCI community in the domain are even more recent [32]. Using 2015 as a starting point, we are sure to capture the uptake in number of contributions in XAI. In addition, we used ACM DL and IEEE Xplore filtering tools to narrow our search to research articles only. In ACM DL, we used the following filter: All Publications/Proceedings/Content type/Research article AND All Publications/Journals/Content type/Research article, therefore excluding surveys, tutorials, introductions, editorials, newsletters, books, magazines, reports, encyclopedias, short papers, extended abstracts, posters, and other non-archival content. In IEEE Xplore, we used the filters Conferences and Journals, leaving out early access articles, magazines, books and standards. This step allowed us to make a first sorting of the non-archived articles, and facilitate the following phase of manual screening. For each record, the article title, authors, publication venue, and publication year were exported to an Excel spreadsheet. Below is the search query used (the wildcards * denote where we have retrieved the plurals and term variants):

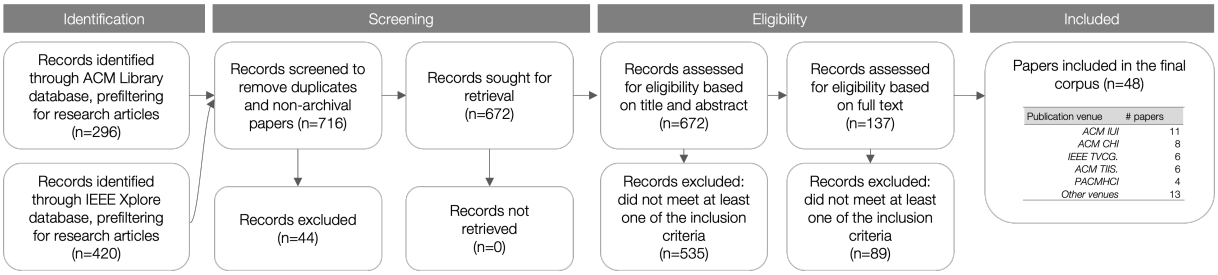


Figure 1: PRISMA flow diagram adapted from Page *et al.* [87] giving an overview of the PRISMA 2020 survey guidelines, used for the search and selection phases of our scoping review.

AI systems => Abstract: (AI, artificial intelligence, machine learning, algorithm*) AND
Explainability => Abstract: (explainab*, explanation*, intelligib*, interpretab*, transparen*, XAI) AND
User studies => Abstract: (participant*, human-subject*, human evaluation*, human experiment*, user-stud*) AND
Date => 2015 or after AND
Journal or conference article => Non-archival records prefiltered out.

3.2.2 Screening. One author deleted 44 records that were either duplicates or non-archival records that remained after the database filtering (primarily workshop entries and student consortia). This step resulted in a total corpus of 672 unique papers.

3.2.3 Eligibility evaluation. The remaining records were randomly assigned to three of the authors, who performed a two-phase eligibility assessment: a first one based on the title and abstract and a second, more in-depth one based on the full text. The first phase was primarily concerned with excluding recordings that were not focused on XAI (IC1, IC2), that did not include a human-AI interaction (IC3), or that were a secondary study (IC7). The second phase consisted of verifying IC4, IC5, and IC6, since full-text viewing was required to assess these criteria. The inclusion criteria were the following:

IC1 *XAI focus.* The paper's contribution is in the XAI field;
IC2 *XAI system.* The paper shows an implementation of an XAI systems;
IC3 *Human-AI interaction.* The paper is in the field of human-AI interaction (works in human-robot interactions are excluded);
IC4 *User-based evaluation.* The paper presents an evaluation of its explainability approach using human-grounded evaluation [32];
IC5 *Human-computer interface.* The paper describes the interface that was presented to the human users evaluating the XAI system;
IC6 *Interactivity.* The explainability approach presented in the paper is interactive, meaning the user can interact *with the explanation* (requiring another interaction than that with the interface to perform a specific task)[1];
IC7 *Primary study.* The paper is not a review nor a position paper.

After the three reviewing authors had completed the eligibility phase, an external reviewer was asked to apply the above criteria to a subset of 67 articles randomly selected from the base of 672 papers, representing 10% of the papers. Inter-rater reliability was 92%, and the remaining disagreements involved mostly cases in which the external reviewer included the articles when the authors did not. However, we believe that the extra step of reviewing the full text in detail is what justified the exclusion of the items that the external reviewer included.

One of the articles included in our corpus [41] was an analysis of an external primary study that did not match our keywords because it did not mention explainability-related terms in the abstract, but it met our inclusion criteria. We therefore replaced the secondary study with the primary study [127].

Eventually, 48 papers met the inclusion criteria and were included in the final corpus.

3.3 Analysis and coding book

3.3.1 Analysis process. The synthesis methodology we used in this review is an emerging synthesis [99], more specifically a narrative account of included studies, as is usually the case in scoping reviews [8]. To support this analysis, we use a concept matrix and a charting approach to provide basic numerical summaries of the extent, nature and distribution of the studies included in the review.

Following Webster and Watson [123], we created a concept matrix for the analysis of the interactivity landscape in the XAI field. The matrix is organized into four dimensions, whether the concepts relate to the *context* of the explanation, its *content*, its *communication*, or its user-based *evaluation*. Three authors independently coded and classified the articles included in the final corpus. For the dimensions context and content, the categories used for coding were predefined. In the communication dimension, only the concept of "representation" had a set of predefined categories. With respect to the type of interactivity, the different categories were intentionally not preset in advance and each of the three coders created their own categories after encountering an interactive explanation implementation. We did this because our goal was to create new categories that matched the range of interactivity types provided by the corpus. The authors then reviewed the resulting categories and discussed how to reconcile them into a taxonomy of interactivity types adapted from well-known existing ones [105, 128]. A similar

[1]Some examples of papers excluded because of IC6 are [11, 17, 31], which present static explanations to end-users, although the user interface to perform a downstream task may be interactive.

Table 1: Codebook used to retrieve information from the corpus with four dimensions: [explanation] *context, content, communication* and *evaluation* and their corresponding sub-dimension. Each paper was assigned a specific code for each sub-dimension. Each paper's sub-dimension may be described by multiple codes. The "Reference" column indicates references from which the coding of the sub-dimension was inspired and adapted from. "NA" indicates the coding was created by the authors.

Dimension	Code	Reference
Context		
Domain	Law and Civic, Healthcare, Business and Finance, Education, Leisure, Artificial, Generic, Other.	[66]
Audience	Domain experts, AI experts/Data scientists, Non-expert, Other.	[66]
Data type	Image, Video, Audio, Tabular, Natural language, Sequential data.	NA
Content		
XAI focus	Raw Data, Output, Model Limitations, Model Confidence How?, Why?, Why not?, How to?, What if?, What's the difference with?, Context.	[70, 71, 113]
XAI method	Local Feature Contribution, Decision Rules, Sensitivity Analysis and Partial Dependence Plot, Example-based, Saliency mask, Concept-based, Surrogate model, Counterfactual, Wizard of Oz.	[66]
Communication		
Interactivity	Clarify, Arrange, Filter/focus, Reconfigure, Simulate, Compare, Progress, Answer, Ask.	[105, 128]
Representation	Chart, Table, Text, Rules, Directly on the data structure, Other.	NA
Evaluation		
Comparison / baseline	No explanation, Static explanation, Other, No baseline.	NA
Evaluation measure	Perceived usability, Perceived usefulness, Understanding, Perceived explanation length/quantity, Time spent interacting with XAI system, Trust, Cognitive load, Performance at task, Learning, Predicted accuracy, Perceived control, Perceived fairness, Perceived transparency, User skepticism, Other. *Only for evaluations using static or no explanation as a baseline, the following codes applied for each evaluation measure:* Higher than, Same as, Lower than [the baseline], Other.	NA

approach was taken for the evaluation portion of the matrix. As new types of evaluations were found, new categories were created. We grouped together concepts that were very similar (such as *explanation utility* and *explanation usefulness*). Finally, evaluations that were used only once in the corpus were regrouped in the "other" category of the matrix. The authors discussed and shared the definition of the notions during several meetings. One author reviewed all the papers and corresponding codings to check the consistency of the two other reviewers' coding with their own, and subsequently consolidated the matrix. Below we detail the different concepts we have analyzed in each dimension.

3.3.2 Context. We retrieved the environment in which the explanations for each item were designed: domain, audience, and data type. The domain and audience categories are adapted from those found by [66] in its survey of AI-assisted decision making tasks. This allows us to see if the interactive explanations are well distributed across these contextual concepts.

3.3.3 Content. To analyze the content of the explanation, we searched for the explanation *focus*, which described the type of information that was provided to the user, and the explainability method used to extract it. The list of explanation focus points was adapted from Lim and Dey's [71], Liao *et al.*'s [70] and from Sun *et al.*'s [113] classifications of user questions in XAI. The categories of the explainability method were adapted from [66].

3.3.4 Communication. Communication refers to the form in which the explanation was provided to the user, including the type of interaction used and the type of visual representation of the explanation. The categories of interactivity are described in more detail in Section 4. The categories of representation were kept general as they were not the focus of this article.

3.3.5 Evaluation. One of the main challenges in XAI is how to measure the quality of an explanation [25]. User-based methods have been an increasingly adopted approach following calls such as Doshi-Velez's [32] to take user perspective into account instead of just technical constraints. While "human-grounded" evaluations may have drawbacks such as sampling bias or change blindness [107], they do inform how end users understand, perceive, and use explanations. This approach also has the advantage that standard questionnaires are shared by researchers to measure concepts such as trust (using the McKnight framework), satisfaction, understanding, cognitive load (using NASA-TLX), etc. We also retrieved the baselines (no evaluation, static evaluation, other explanation, etc.) used to evaluate the presented explanation in each empirical study. This makes it possible to compare the results of multiple studies

and to get an overview of assessments of interactive explanations. For each evaluation in the corpus that used either static or no explanation as a baseline, we reported the results according to four categories: higher than, same as, lower than the baseline, or "other", which referred to more nuanced results dependent on other external factors, or to evaluations that did not rely on a defined baseline.

4 INTERACTIVITY TYPES IN EXPLAINABILITY

Let us now describe the categories of interactivity in XAI that we have identified in our corpus. We took inspiration from other existing taxonomies of interactivity [105, 128] to define these categories. This section addresses our RQ1 and RQ2.

Nine different categories of interactivity in XAI emerged from our analysis. Following Yi *et al.* [128] and Roth and Mattis [95], we formulated the categories so that they express interaction actions that correspond to user intents. We adapted some categories from Sims [105] and Yi *et al.* [128]. However, contrarily to Yi *et al.*'s taxonomy, the object of the interaction are explanations instead of datapoints. Explanations are larger constructs encompassing a visual representation, an input data range, an AI model's configuration (dataset, model type and parameters) and an explainability technique.

In addition to the categorisation of interaction types, we organized the taxonomy into three different groups corresponding to the type of support they provide for the human cognitive process of explaining.

This higher-level categorization is based on Miller's review [77] of social science findings on properties of human explanations. Miller points out that explanations are selective, contrastive, and social. First, explanations are selective as they involve only a few causes in a large chain of causal events. Only a few causes address the explainee's question and are thus relevant. Then, explanations are contrastive as they are thought in contrast to a specific foil. People's questions are almost always "why" questions implying a foil: "why did P happened *and not Q*?" To assess the plausibility of a factor as a cause of an event, people then need to perform mental mutations, *i.e.* to cancel a factor which might have led to P and see if Q happens, or to consider situations where Q happened instead of P. This mental process is called the *mutability* of events and allows the formation of contrastive explanations. Finally, explanations are social because they are best understood in a conversation. The structure of the dialogue allows people to get specific answers to their "why" questions and corresponding foils, to ask follow-up questions and progressively fill the gaps in their knowledge.

Our proposed interactivity groups reflect the degree to which the interactive features enable these explanatory properties—selective, mutable, social. The three categories are: **select** (interactive features facilitate the selection of causes and the formulation of hypotheses), **mutate** (interactive features allow users to compare or simulate different configurations of the AI's inputs, outputs or parameters), and **dialogue with** (interactivity allows users to engage in a conversation with the XAI system). The resulting interactivity taxonomy is outlined in Table 2.

Below we describe in detail the nine different categories of interactive explanations, as well as three levels of interaction into which they fall.

4.1 Select

The user may be able to select[2] the information they wish to see by clicking on hyperlinks to display explanations on demand, by configuring the explanation space, or by filtering the explanation conditionally on an input metric. These interactions can help users formulate hypotheses and actively search for factors that may lead to causal explanations. As such, they enable explanations to be "selective".

4.1.1 Clarify. This subset of interaction capabilities enables the user to make on demand information appear, whether by clicking on or by brushing explanation components. In this approach, the user actively seeks answers to their questions, controlling what explanation to display and when it should be displayed. This set of interaction techniques is close to Yi *et al.*'s "elaborate" category [128]. The analysis of our corpus revealed three main ways for a user to get clarification on something. First, users can navigate through a menu so as to choose the themes they want to know more about. Sims [105] refers to this interaction technique as "hierarchical interactivity". Anik and Bunt [7] is an example of this interactivity type. Second, explanations can be displayed after a user clicks on a link, following Sims' [105] "hyperlinked interactivity". One example is Sovrano and Vitali's [109] explanation system in which the user can click on a concept to get more information about it. With each click, a new window with an explanation appears, itself providing other hyperlinks about the notions used in the explanation. Finally, tooltips are convenient interaction techniques to provide clarifications and additional details on a visualisation in a non-overwhelming way [3, 51, 103, 104]. *Clarify* interactions also allow the explanation interface to be less overwhelming at first glance by disclosing explanations progressively. In a study on the progressive disclosure of explanations, Springer and Whittaker [111] note that "because transparency is provided 'on demand' this removes confusions and inefficiencies arising from spurious, unwanted explanations, and adjusts explanations to the users' requirements." They also observe that this on demand disclosure approach is able to adapt to the different reactions and expectations of each individual user.

4.1.2 Arrange. Arrange interaction techniques provide the user with the ability to organize the explanation space as desired by hiding or collapsing explanations and selecting the type of explanation to be displayed [65]. It is similar to the "rearrange" category in Yi *et al.* [128]. Instead of interacting for more information, (which corresponds to the *Clarify* category), here the user's goal is to configure the explanation space following their preferences. For example, in [73], users can increase or decrease the number of highlighted words in the saliency-based explanation. In [24], the user can chose

[2]A parallel can be drawn here with the "select" category from Yi *et al.* for the Infovis domain, which is defined as "marking something as interesting". Assuming we view this level of interaction as "marking an explanation as interesting", we found, however, several subcategories of interaction types that could be used to support this. This justifies why we refer to it as a whole interaction level instead of just one category.

Table 2: Two-level taxonomy of interactivity techniques in XAI, including a first level reflecting the type of support interaction techniques provide to the cognitive process of explaining, a second task-oriented level, and corresponding definitions.

Cognitive support	Category	Definition
Select	*Clarify*	Give additional information/explanations on demand
	Arrange	Choose and organize the explanation type(s), parameters and visual representation(s).
	Filter/focus	Filter the explanation according to an input/input metric.
Mutate	*Reconfigure*	Change the dataset, the AI model, AI model parameters and show me the corresponding prediction and explanations.
	Simulate	Change the inputs, the output or the dataset distribution and show me the corresponding prediction and explanations.
	Compare	Show me explanations of related or selected data inputs or outputs.
Dialogue with	*Progress*	Guide user through an explanation sequence.
	Answer	Give feedback, edit explanation components.
	Ask	Ask iterative questions and receive answers following a dialogue structure.

the surrogate model used in the explanation along with the other parameters for that model.

4.1.3 Filter/focus. Inspired by Yi *et al.*'s "filter" category, the *Filter/focus* class regroups controls that let the user zoom either on specific inputs of the AI model or subgroups in the the training or testing dataset. The user can therefore focus their attention on the explanation built from a restricted input space. The explanation interface presented in Jacobs *et al.* [48] is an example of a *Filter/focus* interaction technique where users (doctors) can filter explanations based on the presence of a specific symptom. In Cheng *et al.* [22], users can create and delete subgroups in the model's input data to see the corresponding explanations for each subgroup. VBridge [20] and ExplainExplore [24] provide the ability for users to select a subset of features to be used in an explanation. We also put in the *Filter/focus* class sorting functions, such as the one in Gamut [47] which lets the user sort input features according to several feature metrics.

4.2 Mutate

Interactive explanations can allow the user to "mutate" causes, *i.e.* to test their hypotheses by simulating or comparing different situations. The resulting explanations are cumulatively selective and contrastive.

4.2.1 Reconfigure. This category includes a set of interactions that offer the possibility to modify the parameters of the AI model such as the dataset, the model type or the model parameters in order to observe changes on the explanation. Users may want to evaluate the impact of these factors on the model's prediction and corresponding explanation to make sense of how the model works. This is especially true when explainability is used to assess the fairness of the model such as in [127] or [68]. The Silva explanation interface [127], similarly to IBM's AIF360 tool [14], allows the user to modify dataset attributes and sensitive inputs to see how it affects specified fairness measures. Various explanation components, including causal graphs and measures of feature importance, change based on the user's chosen dataset settings.

4.2.2 Simulate (inputs). Interactive explanations can be useful for users to test how changes in inputs affects local explanations and the outcome of the model. Understanding of a model then comes not only from static information about the AI algorithm, but also from the learning experience provided by repeated simulations of the model. Interactions in the *Simulate* category refer to mutations of the inputs of the AI model. Many articles in our corpus (18/47) have integrated this interactive feature, reflecting an appreciation of the XAI community for "learning by doing" [97]. The simulation functionality is usually activated by sliders or drop-down lists and gives the user a local understanding of the model's behavior. Examples can be found in [4, 73, 81, 103].

4.2.3 Compare. This category gathers interaction techniques that are used to compare either (1) explanations for different inputs or group of inputs or (2) explanations for different predictions. In the first case, the user can select the inputs or input groups to compare so as to analyze differences in the explanation. Connections, similarities and differences between the selected inputs or outcomes can be highlighted in the comparative explanations. Compare interaction methods would often use parallel coordinates graphs to ease the comparison between explanations. Hohman et al. [47] give an example of an explanation view in which the user can see local explanations for two inputs they selected for analysis. The second case occurs when the AI model predicts several possible outcomes with varying levels of confidence. The user then usually wants to compare the explanations for each of the probable outcomes to assess their likelihood. Dodge *et al.*'s [30] and Jin *et al.*'s [51] systems are examples of this type of outcome comparison. In Dodge *et al.* [30], the user can tap on a game board (representing a game situation) to see its corresponding chance of winning and how it compares to the chance of winning from other game boards. In CarePre [51], doctors are users, and can explore in detail the records of a patient, as well as compare it with similar patients; their focus

is on sequences of "events" (a patient enters the medical facility, a scan is performed, etc.). This allows the user to detect similar paths, and adapt treatment accordingly. This interaction class is inspired from Yi *et al.*'s "connect" category.

4.3 Dialogue with

Interactivity can support the user in engaging in a dialogue-like structure. Information about the AI model is then given progressively and/or iteratively. The user could ask the system a question or give it feedback. These "dialogic" explanations are in line with the properties expressed by Miller for human-like explanations. However, there may be different degrees in which explanations are truly social, depending on the range of questions a system can actually answer.

4.3.1 Progress. The *Progress* interaction style is inspired by Sim's "linear interactivity" through which "the user is able to move forward or backward in a pre-determined sequence of instruction materials". The explanation is designed in several steps, and the user can click "next" or "previous" to navigate through the explanation displays. It is generally progressive, with basic information provided in the first few pages and more in-depth information presented in subsequent sections. This style of interactivity is reactive [105] and does not provide specific feedback to the user but instead lets them walk through the explanation at their own pace. The user can only control when the explanation is provided. The *Progress* interaction style can be seen as the lowest level of "dialogic" explanations. It does not enable the user to ask nor answer questions but it follows some of the rules of a conversation [40] by providing sparse information progressively (maxim of quantity), and by predefining user questions that need to appear in the explanation guide (maxim of quality). The "next" and "previous" commands can be considered as the users' options to punctuate the conversation (compared to saying "ok tell me more" or "wait, what did you say").

4.3.2 Answer. While information flow in interactive XAI systems goes primarily from the machine to the user, like in Infovis systems [128], it can also be reversed, with users providing the system with feedback, corrections or information about the state of their mental models. These interactions can serve to increase users cognitive engagement (and activate their "System 2" [52]) by challenging users. For example, in [75], users (in this case children) are asked to click on the part of the image that they think had the most impact on the algorithm's prediction. This interaction type can also serve to improve the AI system by building on human feedback. Examples are [50, 104] in which users are asked to improve the semantic meaning of the concepts learned by the algorithms, [21, 42] in which users can create or edit explanations—such as adding a new rule or correcting one, [121] in which users can indicate to the system their personal preferences about model interpretability, or [38, 39, 43, 110].

4.3.3 Ask. In Miller's view [77], the ultimate level of interaction is a conversation where the user can ask the AI system anything they want. We can therefore view the *Ask* interactivity as the higher end of the interactivity scale for XAI. The conversational XAI research line has made some progress in achieving such interactivity. For instance, [43, 44] present logical dialogue maps to deliver explanations that answer users' questions. The challenge is to cover as wide a range of questions as possible. Note that this "dialogic" interaction between user and machine does not necessarily have to take place through natural language. As Miller stated [77], we could imagine an XAI system that answers the user's questions with images or other communication means. An illustration of this can be found in [55], where the user submits a query such as "create a graph showing the predicted trend" and the XAI system responds with the desired graph.

5 ANALYSIS

In this section, we present a qualitative analysis based on our conceptual matrix to address our RQ2 (Section 5.1), RQ3 (Section 5.2), and RQ4 (Section 5.3).

5.1 In what context, with what content, and in what form were the interactive explanations presented to users?

5.1.1 Context. The work in our corpus is well distributed across the different domain categories constituted by [66] (*cf.* Figure 4 in the Appendix). Notably, the corpus reflects a large number of studies (32/48 papers) implemented in real-world applications rather than in artificial or generic domains. Healthcare stands out as one of the most studied domains in the corpus.

Some work [16, 22] expressed concern that too few studies focused on making explanations understandable to novices and that most current XAI techniques were only comprehensible to AI-educated users. Cheng *et al.* [22] also argues that the majority of studies providing explanations to novices have been conducted in the context of generic tasks [66], *i.e.* computer science problems, and are therefore not generalizable to real-world applications. In contrast to the first concern, we found that the majority of articles included in the corpus (27/48) were aimed at a general audience of non-expert users. This at least reflects an awareness of the field to design explanations with this user group in mind. In addition, 15/27 of these studies are in real-world application areas, including areas that may be considered sensitive—4 in legal and civil, 2 in healthcare, and 3 in business and finance. However, it is possible that the empirical studies included in our corpus targeted non-expert users for practical reasons, such as to solicit platform workers like those on Amazon MTurk [22, 38, 42, 44, 93, 94, 98, 124]. Nevertheless, some of these studies are primarily aimed at making the XAI systems more transparent and more accessible to a non-expert audience [7, 111, 116, 119, 127].

Regarding the data type used in our corpus, tabular and text data are predominant (79% of the studied papers). This points to an opportunity for the XAI field to empirically study interactive explanations using audio (only one paper discussed audio data [7]), images, and video data.

5.1.2 Content. The interactive explanations in the corpus focused heavily on the "why?" user question recurring 37 times, and which can be answered by local feature explanations, the most commonly used explanation method in the corpus (26/48). We can see in Figure 2a how some interaction techniques were favored for specific

types of user question. For example, quite logically, explanations addressing "what is the difference with?" were implemented with *Compare*, but also frequently with *Filter/focus* interactions. Context and raw data can be elaborated through *Clarify* interaction. "How to?" and "What if?" were facilitated through *Simulate* interactions. Model limitations were rarely presented in the studies (only twice). But perhaps a bigger opportunity for interactive explanations is the small numbers of papers addressing "how to?" questions. One example is [94] in which the user can change input "concept features" to see the adjusted output in real time and better understand the meaning of each "concept feature". However, we found only two studies enabling direct interventions on the model output [30, 51]. Such interventions (which would fall in the *Simulate* category *cf.* Table 2) could help the user characterize what kinds of contexts and situations are emblematic of a particular outcome, thereby addressing "how to?" questions. In addition, concept-based explanations, which are considered promising in the field for their human comprehensibility, were rarely used in the corpus [56, 59].

5.1.3 Communication. The most used interaction techniques were *Clarify* and *Simulate*. These were frequently combined with *compare, Filter/focus* and *Arrange* as illustrated in Figure 2b. The techniques *Progress* and *Ask* were used in only three and four studies respectively, illustrating a trend in the field of interactive XAI towards complex, Infovis-type XAI interfaces rather than simpler step-by-step or dialog box interfaces. The matrix in Figure 2b shows this clear cut between the "Select" and "Mutate" interaction groups on the one hand, and the "Dialogue with" group on the other. The interactions techniques in the first two groups are frequently combined with each other, while the interaction styles in the latter group are less frequently used. In addition, these more "social" interactions were rarely combined with other interactions from the "Mutate" or "Select" levels. In particular, *Progress* was never used in combination with other "Mutate" or "Select" interaction categories, as shown in Figure 2b. It would be interesting for future research to explore combining these as a way to take advantage of the social nature of "progress" explanations while giving greater control to the user with selections and mutations.

The representations used for the interactive explanations were primarily charts and texts. As shown in Figure 2c, tables were useful to support *Filter/focus* and *Compare* interactions. Textual explanations often came with *Clarify* interactions. Rules, although not appearing frequently in the corpus (5 times), where used to support *Clarify* and *Answer* interactions. Indeed, rules are easy objects for users to modify, create or delete, as exemplified in [42, 43, 78].

5.2 How were interactive explanations evaluated?

To address our RQ4, we describe below the different user-based evaluation methods and measures that have been used in our corpus to evaluate XAI systems and explanations. Below we provide brief descriptions of the measures and highlight trends and challenges in evaluating interactive explanations.

5.2.1 Few controlled experiments. Few empirical studies supported a cross-sectional analysis of results on interactive XAI by using a static explanation as a baseline. Most papers (20/48) did not use any control condition (*cf.* Figure 4 in the Appendix). Even if the measures in these articles are sometimes quantitative as in [44] where the authors measured different constructs (system efficiency, transparency...) on Likert scales from 1 to 5 points, these results are hard to interpret in comparison with the rest of the XAI literature.

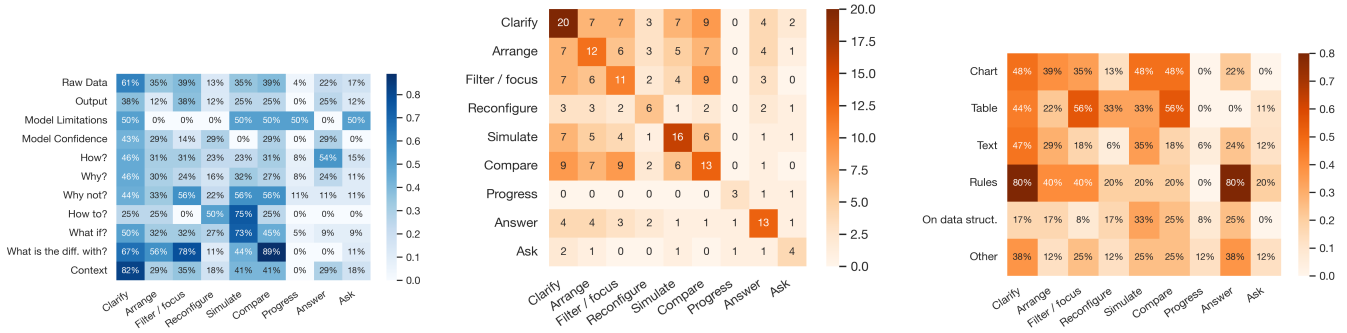
Nine of the 48 articles in our corpus compared interactive and static explanations through between-subject experiments. These comparisons were very informative for analyzing the added value of interactivity in XAI. We provide in Section 5.3 a qualitative analysis of the added value of interactive explanations based on this work. To a lesser extent, comparisons between interactive explanations and no explanation (13/48 items) are also useful for understanding the benefit of interactive explanations. We also leveraged this body of work in Section 5.3. Other context-specific comparisons were made between an interactive explanation and other explanation types [42, 98, 115, 124], other interactive systems [89, 127], other AI models [94], other interactivity types [38] or random baselines [50], among others. Some of these user-based evaluations were within-subject experiments [30, 35, 111].

Much of the work that did not use a baseline provided valuable qualitative assessments instead. This research often employs usage scenario (or "use cases") to study users' reactions to XAI systems in realistic settings [20, 50, 65, 78]. These qualitative insights often focused on capturing the user's perceived ease of use and/or usefulness of the XAI system (16/20 papers).

5.2.2 A wide toolbox. We identified 19 different metrics to evaluate XAI systems with users from our corpus. Fourteen of them were used twice or more: perceived usability, perceived usefulness, understanding, perceived explanation length/quantity, time, trust, cognitive load, performance at task, learning, predicted accuracy, perceived control, perceived fairness, perceived transparency and overtrust (*cf.* Figure 4 in the Appendix). Other measures were used such as perceived feedback quality and difficulty [42], explanation persuasiveness and sufficiency [44], number of interactions (clicks, etc.) with the explanations [84] and naturalness and humanness of the explanations [91]. Table 3 provides the definitions used for each of these metrics.

We recognized four of the five user-based measures for evaluating XAI systems outlined in [46]: user satisfaction, understanding, trust (and overtrust) and human-XAI performance. Indeed, none of the papers in our corpus measured participants' curiosity, highlighting a gap in the literature for making XAI systems more engaging through users' feedback. However, we actually found more than five types of human-based metrics. Measures of the propensity of XAI systems to enhance learning, perceived transparency and fairness, humanness and naturalness of explanations, or cognitive workload, provide additional nuances to the XAI researchers' toolbox.

5.2.3 The many shades of user satisfaction. User satisfaction was the most frequently used measure in the corpus. However, we found many nuances of this concept. Some assessed whether users liked the systems [50, 57], and/or found them useful [51, 55, 109], helpful [48, 127], effective [44] and/or easy to use [65, 116], or preferred the explanation or explanation system over another. In order to capture some of these nuances while keeping the papers coding manageable, we divided user satisfaction into two main clusters:



(a) Percentage of studies focusing on a type of user question per interaction category (*e.g.* 4% of studies focusing on raw data use the *Progress* interaction). One study can feature multiple interaction categories.

(b) Frequency of the interaction categories used in the corpus and frequency of their combinations (*e.g.* 20 studies use the *Clarify* interaction and 7 use both *Clarify* and *Arrange*).

(c) Percentage of studies using an explanation representation per interaction category (*e.g.* 48% of studies using charts use the *Clarify* interaction).

Figure 2: Summaries of the corpus through different concept matrices.

ease of use (*i.e.*, perceived usability) and perceived usefulness of the XAI system.

Some articles already made distinctions between these two constructs [51, 116], but others did not, especially when using questionnaires such as the Explanation Satisfaction Scale [46], which incorporates both usability and usefulness concepts [16, 42]. When this was the case, we reported the measure under both "usability" and "usefulness". Under the "perceived usability" construct, we included measures of usability, ease of use, likeability, *i.e.* whether users expressed that they liked the interactive explanation (or the XAI system)—typically through a one-item questionnaire [42] or through a qualitative think-aloud study [51],—and user preference, *i.e.* whether users preferred the system to a given baseline. Questionnaires such as the Post-Scenario Questionnaire [69] or the User Engagement Scale [86] were often used to measure usability. In the concept of usefulness, we reported the accounts of "usefulnes" and "perceived effectiveness", the latter being assessed through Tintarev's questionnaire [44, 117, 119].

5.2.4 Joint use of subjective and objective measures. Many self-reported measures have an objective equivalent, and the papers in our corpus have taken advantage of this. This was the case for understanding, trust and cognitive load.

Understanding was most often measured subjectively by asking participants if they understood the system [16, 23]. However, some also assessed understanding objectively by asking carefully designed, often context-specific questions [16, 22, 78, 91]. Predicted accuracy, referring to the ability of users to predict what the system will output given certain entries, has been measured in [23, 85, 111] and could be considered, as some argue [23], as an objective understanding of the system.

Participants' trust in the system or explanations was mostly assessed subjectively, by asking people to report their confidence in the XAI tool. McKnight's framework was used in three studies [38, 44, 124]. Other papers referred to Tintarev's [117] measures of trust [44, 119, 124]. [44] also used items from Kouki *et al.* [61] to measure trust related to explanations rather than to the system. However, trust was also measured objectively, by observing users' ability to reject an incorrect AI suggestion [18, 57, 73, 93]. We referred to this measure as "overtrust", but [57] framed it more positively as "user skepticism", while others have called it "human-AI agreement" [73].

Users' cognitive workload when interacting with XAI systems was reported in five studies. It was measured by the NASA-TLX workload index, or a subset of its items. Closely related to cognitive load are estimates of the time spent on the XAI system or explanation, and the perceived length and/or complexity of the explanation. The former is an objective, quantitative estimate, while the latter is a self-reported measure [18, 62, 116].

The quality of self-reported measures can sometimes fall short of researchers' expectations, as some [30, 84, 122] argue. Objective measures of understanding, trust and cognitive load may offer more reliable observations, even though at present, their measures are less standardized and more context-specific, making results more difficult to compare across different studies. Dodge *et al.* [30] notably proposed "the ranking task" as an alternative to self-reported measures.

5.2.5 Task performance as the new benchmark. Some work [11, 17] advance that subjective measures could be misleading to properly assess the added value of explanations. Buçinca *et al.* [17] found that an increase in user satisfaction did not necessarily lead to improved performance, if not the opposite. Instead, Buçinca argues, measuring task performance should be the standard benchmark as it comes down to directly evaluating XAI systems against what they were designed for: increasing humans' autonomy and complementarity with AI. While XAI may serve other purposes, such as increasing user confidence and understanding, measuring task performance has the advantage of being a metric that is both objective and easily quantifiable. In fact, many empirical studies in the corpus have adopted it (21/48). Some articles also measured other constructs related to the task at hand, such as task complexity or time spent performing the task [94].

Table 3: Evaluation concepts used twice or more in the corpus with the corresponding definitions used in this review and the corresponding evaluation methods used in the surveyed papers.

Evaluation concept	Definition	Main evaluation methods
Perceived usability	User's perception of how easy to use the explanation user interface is.	Adapted question items from Explanation Satisfaction Scale [46], Post-Scenario Questionnaire [69] or the User Engagement Scale [86]; qualitative think-aloud study [51].
Perceived usefulness	User's perception of how useful, effective or helpful the XAI system is for achieving their goals.	Question items from Tintarev's questionnaire [117], Explanation Satisfaction Scale [46] or [120]; qualitative think-aloud study [127].
Understanding	The extent to which the user understands a model or its explanations.	"Objective understanding": Likert-type, context-specific questionnaires [16, 22, 78, 91], "Subjective understanding": qualitative think aloud or free-text analyses, *e.g.* [16, 23].
Perceived explanation length/quantity	User's perception of the length or quantity of the explanation, often used as proxies for the complexity of the explanation.	Direct questions about the quantity, length, or complexity of the explanation, *e.g.* [18, 62, 116].
Time	The time spent by the user interacting with the XAI system to perform a task.	Direct measure of the interaction time, *e.g.* [94].
Trust	User's willingness to depend on an XAI system because of the characteristics of the system [74, 96].	Question items from McKnight's framework [74], Tintarev's questionnaire [117] or Kouki *et al.* [61]'s measure of trust towards explanations.
Cognitive load	The amount of working memory resources used by the user while interacting with the XAI system [79].	NASA-TLX workload index.
Performance at task	The performance of the human+XAI team in performing a specific task.	Measured through case-by-case metrics adapted to a context-specific task, *e.g.* [18, 30, 35].
Learning	How well explanations and/or XAI systems help users learn about a specific topic.	Context-specific questions usually defined by the authors themselves about a topic. See examples for learning about gender bias ([75]) or self-care awareness ([119]).
Predicted accuracy	User's ability to correctly anticipate the AI's behavior.	Number of correct guesses of the AI's prediction by the user [23, 85, 111].
Perceived control	User's perception of their control over the XAI system.	Adapted question items from the Knijnenburg et al. [58] framework.
Perceived fairness	The extent to which users perceive the XAI system to be fair and transparent.	Fairness questionnaires from [15] or Lee *et al.* [68].
Perceived transparency	User's perceived understanding of the recommendation rationale	Adapted question items from Millecamp *et al.* [76] or Tintarev [117] frameworks.
Overtrust	User's ability to reject an incorrect AI suggestion.	Precision and/or recall in correct rejections or acceptances of a prediction, *e.g.* [18, 57, 73, 93].

5.2.6 Less frequent goal-specific metrics. Evaluation measures are chosen in relation to the purpose that explanation serve. For example, Lee *et al.* [68] and Anik *et al.* [7] aimed at increasing public transparency and perceived fairness of an AI system. Therefore, Anik *et al.* used the questionnaire from [15] to assess users' perception of the fairness of the system and Lee *et al.* relied on their own quantitative metrics by asking participants to indicate on a Likert scale their agreement with the sentences "My assignment is fair", "This participant's assignment is fair", or "The overall group outcome was fair". Similarly, learning was a few times measured as a separate concept from the understanding of the AI model. Measures of "learning" focused on how well XAI explanations and systems helped users learn about a topic such as gender bias ([75]) or self-care awareness ([119]). In conversational interfaces, explanations were evaluated according to their humanness and engagingness [43, 102], to their persuasiveness [44], or their naturalness [91].

5.3 What are the effects of interactive explanations on user-based measures?

Previous research has demonstrated uncertainty about the benefits, if any, of interactivity in XAI. While theoretical work in education and psychology outline the benefits of interaction for explanation

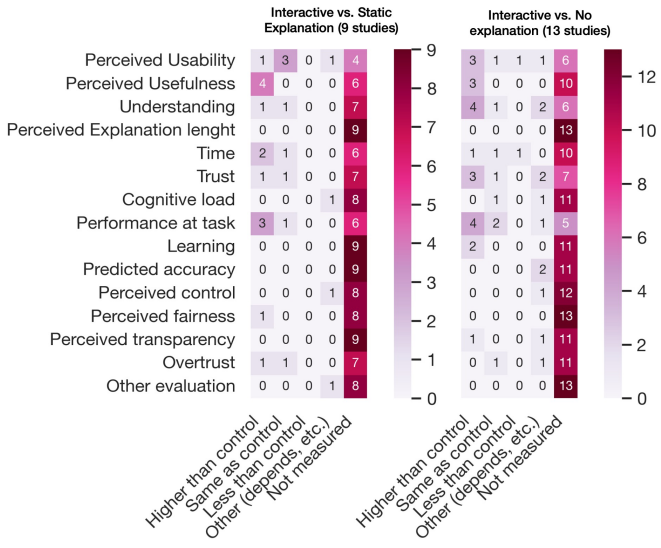


Figure 3: Left: Summary of the effects of interactive explanations compared to static ones, against various user-based metrics, based on 9 different studies. Right: Summary of the effects of interactive explanations compared no explanation as a baseline, extracted from 13 different empirical studies.

and learning [77, 97], empirical results do not always align with these statements. In [73] for example, they find that interactivity could increase human biases and over-reliance on AI. In an attempt to resolve this ambiguity, we present a summary of empirical evaluations of interactive XAI below. We base our qualitative findings on the summary presented in Figure 3, and on the qualitative analyses of the effects of interactivity provided in the corpus. The results presented here are not intended to clinically demonstrate evidence for a hypothesis, but rather to identify the types of qualitative results that emerge in the interactive XAI field and help orientate following systematic reviews.

5.3.1 Interactive explanations improve perceived usefulness but not usability. Overall, there appears to be repeated evidence that interactivity does not significantly improve perceived usability [42, 68, 109] compared to static explanations, but it does improve perceived usefulness [16, 18, 38]. However, when compared to a baseline of no explanation, interactive explanations lead to an increase in perceived ease of use [43, 57, 119]. This reinforces the hypothesis that interactivity is not responsible for the improvement in perceived usability, but the presence of explanations is. It is possible that interactivity increases the complexity of the system, but at the same time supports users in their task and exploration of the models. The authors of the Gamut interface [47] state that "interactivity was so fundamental for our participants' understanding of the models, that when we prompted them to comment on interactivity, people could not conceive non-interactive means to answer both their hypotheses and prepared questions". This study illustrates the potential of interactivity in terms of usefulness and as a factor in enabling users to achieve their goals.

5.3.2 Interactive explanations improve performances of the (human+AI) team, sometimes increasing time spent on explanations. Human+AI team performance was found to be improved in [18, 39, 68] with interactive versus static explanations. However, in two other studies [18, 22], the time spent to interact with the explanation system was higher for interactive explanations compared to static ones. The presence of interactive explanations compared to a "no explanation" baseline also improved task performance. These results seem logical, as greater interactivity can help users dive deeper into exploring a model and augment their cognitive engagement in the process. However, increasing the number of interactions with the system, as well as deeper analytical thinking, would understandably take more time. For example, interactivity can be designed to elicit user cognitive engagement such as in [18], which in turn can enhance task performance. Further, Buçinca *et al.* [18] showed that on demand explanations—from the *Clarify* interaction category—could significantly increase the performance of the human+AI team compared to static explanations. However, Naiseh *et al.* [84] demonstrated that an interactive friction-based feature—falling in the *Answer* category—could lead participants to interact significantly more with the system, while having no impact on the time spent using the system.

5.3.3 No clear indication of an interactivity effect on cognitive load or over-reliance. Some concern has been expressed that interactivity could increase users' cognitive load and their over-reliance on AI [73]. We did not find many results to either confirm or refute this. The results for user cognitive load were generally not directly related to explanations alone, but to other external factors, either with the static or no-explanation baseline. Buçinca *et al.* [18] and Ghai *et al.* [38] highlighted the importance of the user's individual need for cognition, knowledge of the task to perform, or of the model used [94]. Qualitative analyses suggest, however, that *Simulate* interactivity techniques can increase users' perceived difficulty of interacting with the system as we detail in the paragraph 5.3.7.

Compared to no explanation, interactive explanations did not lead users to over-rely more on the AI. However, results were mixed for the comparison of interactive explanations to static ones. On the one hand, using *Simulate* interaction techniques, [73] found that interactive explanations could increase users' tendency to blindly trust the AI. On the other hand, [18] found that their on demand interactive features in the *Clarify* style could significantly decrease over-reliance. The interactivity type therefore seems to be instrumental in the development of over-reliance.

5.3.4 Higher perceived control leads to greater perceived fairness, perceived transparency, and (less clearly) trust. A participant in [127] said "I want to know why it is biased, not have the machine tell me why". This highlights the power of user controls and interactivity to drive trust and support users' autonomous exploration of the AI model. Lee *et al.* [68] confirmed this with quantitative evidence, finding that *Reconfigure* interactions significantly improved perceived fairness. The authors mentioned that the *Answer* interaction—here participants could correct the algorithmic allocation—caused users to perceive the model as fairer.

We did not find a substantial trend in the effect of interactivity on trust in the quantitative studies in the corpus. As indicated by the right side of Figure 3, the results in [55] and [22] do not converge.

Some studies described the link between trust and external factors such as users' prior experience with AI [38] or on users' individual propensity to trust [57].

5.3.5 Unclear role of interactivity on understanding and learning. From Figure 3, it appears clearly that the presence of (interactive) explanations compared to no explanation enhances user understanding of a model. Similarly, learning seems to be persistently enhanced by the presence of interactive explanations [75, 119]. At the same time, user understanding of a model was dependent on other factors, including the order in which users saw weaknesses in the system [85], or the stage of interaction with the system [23], or the type of model that was explored [94]. In addition, [22] found that interactive explanations led to higher objective and subjective understanding of the model compared to a static baseline, but [16] could not find any statistically significant improvement of interactive over static explanations for both objective and subjective understanding. More work is therefore needed to clarify the added value of interactive explanations over static explanations for understanding and learning.

5.3.6 Qualitative evidence of the added-value of a few interaction techniques. Despite the unclear quantitative evidence, the qualitative analysis of the corpus suggests that understanding is facilitated by interactivity. For example, one participant reported that receiving feedback and interacting with the model helped him "learn from my mistakes and expose my misconceptions" [30]. Sevastjanova *et al.* [103], showed that participants appreciated the on demand display of explanations as well as the ability to edit them. Morrison *et al.* [81] emphasized the usability of *Compare* interactive features to support human cognitive processes, finding that "comparison is much easier than classification for a person". Khurana *et al.* [98] demonstrated qualitatively that linear interactivity was perceived as useful. Furthermore, Springer and Whittaker [111] highlight the need for progressive disclosure of model information in order to prevent users from seeing their expectations violated and distrusting the system when it is correct.

5.3.7 Simulate interactions can strain users' memory and time. While interactive explanations of the type *Simulate* have been evaluated positively on many fronts, notably usability, usefulness and understanding, they also seem to take up more time as qualitative analyses in [16, 38] show. Additionally, after using a simulation-based interaction feature, a participant in [50] indicated that: "At the end of the design process, I think my brain is stuck. I do not know what I have specified before. When I want to add a new attribute, I need to go back to check if I have specified it already". This calls for a careful consideration of the natural tendency of people to lose track of previous simulations in the design of *Simulate* interactions. Consistent with this observation, [94] found that user performance in recreating an outcome through perturbations of concept-features degraded as the dimensionality of the concept-features increased. Future research should therefore design simulation explanations taking into account the limitations of people's memory.

5.3.8 Current dialogic explanations lack humanness. In [91], participants rated the naturalness of conversational explanations more harshly than the other measured aspects of the explanations. Also, in [119], participants reported a similar lack of naturalness for the questions that were asked by the system to the user. The authors describe: "our participants felt confused about the questions asked by the [conversational agent] in terms of the sequence, quantity, and relevance." However, in [43] participants indicated they preferred to be able to "recognize when they were talking to a human or to a machine", actually preferring that humanness levels of explanations remain low. This questions the validity of aiming for more "dialogic" explanations that replicate a human-like explanation process. We provide more thoughts on this issue in the following section.

6 DISCUSSION

We discuss below two open issues in interactive XAI. First, interactivity itself needs to be explained to users, adding another layer of complexity to XAI systems. Second, it is unclear whether dialogic/human-like explanations should be considered the ideal form of explanation communication by XAI researchers.

6.1 Interactivity calls for meta explanations

Interactivity itself requires some learning by the user [97]. In addition to learning about the model, users must learn how to use the controls of the interface.

Hepenstal *et al.* [43] observed that participants had many questions about how to use the interface and control it—"Can I click on that ?". With *Answer* interactions, Tsai et al. [119] also found that some participants felt confused by the questions asked by the system. They suggest that it would be helpful to provide additional explanations answering questions like "why does the system ask these questions?", or "how many questions would be asked or needed?" [119]. These observations align with Sun *et al.*'s categorisation of user questions. One of them is called "Control", and is defined as "Questions about options for customizing or specifying preferences for how the model should work". Therefore, interactivity adds a layer of explanation in addition to model explanations.

We can make a parallel with the concept of meta-explanation introduced in [26]. Dazeley *et al.* point to a major issue in XAI research, which is the user's need to know where explanations come from in order to be able to trust the model and its explanations. As the authors put it: "if we cannot trust the agent's original decision, how can we trust the agent's explanation of that decision?". They call "meta-explanations" the explanations about the explanations themselves. Meta-explanations introduce a paradox whereby more explanations calls for more explanations, leading to unsustainable complexity. Similarly, explanations on the control of the interface could lead to cognitive overload and effects such as users ignoring explanations and AI predictions, as described in [119].

Our corpus highlighted diverging results on whether interactivity has an effect on cognitive load. Our analysis highlighted, however, the role of individual factors to drive cognitive workload. There is therefore a need for future research to investigate how to tackle the meta explanation paradox in the context of interactive XAI, and how to find the right level of explanation for each user [18, 26].

6.2 Are dialogic explanations really the grail?

According to Miller [77] and Graaf and Malle [27], people expect explanations to follow the conceptual framework of a social interaction. One reason for this is that people attribute human traits to XAI agents and therefore expect them to follow social conventions [27]. Therefore, a good explanation would be provided through a social conversation. In fact, at least two studies from our corpus provided quantitative evidence that explanations communicated through *Ask* interactions improved perceived usability and understanding.

However, the participants in Hepenstal *et al.*'s study [43] were bothered by the humanness of the XAI agent and preferred to have it made clear that they were not talking to a real person. Instead, they preferred robot-like explanations with "logical and clear responses". Indeed, while explainability should bring trust, anthropomorphism through human-like conversations can diminish trust by giving people the feeling of being manipulated. Hepenstal *et al.* suggest that different evaluation metrics could be applied to assess conversational XAI, such as understanding and bias mitigation, which are more representative of explainability's purpose.

If we take Miller's [77] depicted ideal of an AI agent's explanation[3], perhaps a more important criteria than the social structure of the explanation would be the range of questions the explaining agent is able able to answer. Overall, further theoretical work may be needed to clarify what "social interaction" means, whether it refers to its dialogue structure or to the social rules it abides by, such as Grice's [40] maxims. Future work could also examine the extent to which a "social" interaction with an AI agent can resemble human conversations, or even if this comparison makes sense.

7 LIMITATIONS

One of the main limitations of scoping reviews is that they do not formally appraise the quality of the included studies [8] through the means of, for example, the Cochrane Risk of Bias or other quality assessment tools. While this is compatible with the objectives of this paper—to identify, map and discuss evidence on empirical results in interactive XAI—we remind the reader again of this limitation.

Furthermore, although we applied a standardized methodology to identify articles, it is possible that relevant papers were missed because they were not published in peer-reviewed conferences or journals, because they were not present in the databases we surveyed or because they did not match our keyword search. This was the case for [106], which was published in a workshop and was therefore excluded during the eligibility phase, or for [126] which did not appear in the databases we searched. Indeed, as mentioned earlier, we chose to focus on HCI-oriented databases (ACM DL and IEEE Explore) rather purely AI ones, which may have led us to leave out relevant work in CS-focused venues. Since our interest is in interactivity and user studies, it seemed reasonable to limit ourselves to academic venues in HCI. Other work like [63] and [64] were not included in our study because the authors use the terms "interpreting" or "explanatory" in their title/abstract as references to the "explainability" notion. However, we believe that it would have been difficult to define the verbs interpret or explain and their conjugations as keywords because of their ubiquity. To remedy the limitation of a keyword search for the interactivity dimension, we searched for papers presenting an interactive XAI system in the eligibility phase instead of the identification phase [80]. This enabled us to include papers presenting interactive XAI solutions even though they did not expressed or emphasized in the abstract their contributions to the interactive XAI field.

In addition, we acknowledge that there may be a positive outcome bias [19] in the results on interactivity because we searched published articles. We hope that by highlighting areas of uncertainty where it is unclear whether interactivity has positive or negative effects, this work will encourage others, including publishers, to consider all types of outcomes, including neutral or negative.

Then, although steps were taken to ensure consistency in our coding—including a final review of all the codings by one researcher—the final matrix may reflect each reviewer's own way of thinking.

Finally, it is possible that the summary of the papers' findings in Section 5.2 may not capture the nuance of each context in which the results were found. However, it does provide a high-level, qualitative view of the results of empirical studies, and that was our goal.

8 CONCLUSION

This paper presented a review of the literature on interactive explanations evaluated with human users. We provided a qualitative analysis of 48 papers shedding light on (1) the types of interactivity techniques that have been used so far in XAI, (2) the context in which interactive explanations were implemented, (3) the metrics used to evaluate interactive explanations with human users, and (4) the effects of interactivity on user satisfaction, understanding, trust, performance at task and other user-based metrics.

We provided a classification of XAI-specific interactivity techniques which can serve as a basis for explainability system designers to navigate the interactivity spectrum in XAI.

Our analysis showed that attention has been focused on interactivity that allows for input modification, but less attention has been paid to output perturbations and to more dialogic interactions. Combinations of dialogic interactions with interactions that allow mutation or selection is an under-explored area. The evaluation metrics we observed provide a wide range of ideas for XAI researchers to evaluate their systems against what they were designed for. Finally, we found converging results regarding the effect of interactive explanations on users. The main empirical results we identified were that interactivity increases perceived usefulness and the performance of the human+AI team compared to static explanations, but it does not improve usability. In addition, it increases time spent by users on XAI systems. The empirical studies gathered in our corpus also demonstrated conflicting results on the role that interactivity has on over-reliance, cognitive load, learning and understanding. This highlights grey areas to be addressed in future empirical research. Finally, we hope that this work will help future research to share a common vocabulary on interactive XAI. Also, we hope it will facilitate future systematic reviews to identify best practices in interactive XAI design, as more empirical research is conducted in this area.

[3]Miller presents it as a conversation, not necessarily in natural language, where the user asks a first request and follow-up questions.

ACKNOWLEDGMENTS

This research is sponsored by the Agence Nationale de la Recherche (ANR) through the grant ANR-20-CHIA-0023-01. The authors thank Joshua Brand, Samuel Huron, Jan Gugenheimer, and anonymous reviewers for helpful discussions and comments.

REFERENCES

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

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

[3] Sabbir Ahmad, Andy Bryant, Erica Kleinman, Zhaoqing Teng, Truong-Huy D. Nguyen, and Magy Seif El-Nasr. 2019. Modeling Individual and Team Behavior through Spatio-temporal Analysis. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play (CHI PLAY '19)*. Association for Computing Machinery, New York, NY, USA, 601–612. https://doi.org/10.1145/3311350.3347188

[4] Yongsu Ahn, Muheng Yan, Yu-Ru Lin, Wen-Ting Chung, and Rebecca Hwa. 2022. Tribe or Not? Critical Inspection of Group Differences Using TribalGram. *ACM Transactions on Interactive Intelligent Systems* 12, 1 (March 2022), 5:1–5:34. https://doi.org/10.1145/3484509

[5] R. Amar, J. Eagan, and J. Stasko. 2005. Low-level components of analytic activity in information visualization. In *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005.* IEEE, Minneapolis MN USA, 111–117. https://doi.org/10.1109/INFVIS.2005.1532136 ISSN: 1522-404X.

[6] Geoffrey R. Amthor. 1992. Multimedia in education: an introduction. *Int. Business Mag.* (1992), 32–39.

[7] Ariful Islam Anik and Andrea Bunt. 2021. Data-Centric Explanations: Explaining Training Data of Machine Learning Systems to Promote Transparency. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–13. https://doi.org/10.1145/3411764.3445736

[8] Hilary Arksey and Lisa O'Malley. 2005. Scoping studies: towards a methodological framework. *International Journal of Social Research Methodology* 8, 1 (Feb. 2005), 19–32. https://doi.org/10.1080/1364557032000119616 Publisher: Routledge _eprint: https://doi.org/10.1080/1364557032000119616.

[9] Vijay Arya, Rachel K. E. Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Q. Vera Liao, Ronny Luss, Aleksandra Mojsilović, Sami Mourad, Pablo Pedemonte, Ramya Raghavendra, John Richards, Prasanna Sattigeri, Karthikeyan Shanmugam, Moninder Singh, Kush R. Varshney, Dennis Wei, and Yunfeng Zhang. 2019. One Explanation Does Not Fit All: A Toolkit and Taxonomy of AI Explainability Techniques. http://arxiv.org/abs/1909.03012 arXiv:1909.03012 [cs, stat].

[10] S. Sandra Bae, Clement Zheng, Mary Etta West, Ellen Yi-Luen Do, Samuel Huron, and Danielle Albers Szafir. 2022. Making Data Tangible: A Cross-disciplinary Design Space for Data Physicalization. In *CHI Conference on Human Factors in Computing Systems*. ACM, New Orleans LA USA, 1–18. https://doi.org/10.1145/3491102.3501939

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

[12] Philip Barker. 1994. Designing Interactive Learning. In *Design and Production of Multimedia and Simulation-based Learning Material*, Ton de Jong and Luigi Sarti (Eds.). Springer Netherlands, Dordrecht, 1–30. https://doi.org/10.1007/978-94-011-0942-0_1

[13] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58 (June 2020), 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

[14] R. K. E. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilović, S. Nagar, K. Natesan Ramamurthy, J. Richards, D. Saha, P. Sattigeri, M. Singh, K. R. Varshney, and Y. Zhang. 2019. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development* 63, 4/5 (July 2019), 4:1–4:15. https://doi.org/10.1147/JRD.2019.2942287

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

[16] Clara Bove, Jonathan Aigrain, Marie-Jeanne Lesot, Charles Tijus, and Marcin Detyniecki. 2022. Contextualization and Exploration of Local Feature Importance Explanations to Improve Understanding and Satisfaction of Non-Expert Users. In *27th International Conference on Intelligent User Interfaces*. ACM, Helsinki Finland, 807–819. https://doi.org/10.1145/3490099.3511139

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

[18] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. 2021. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 188:1–188:21. https://doi.org/10.1145/3449287

[19] Michael L. Callaham, Robert L. Wears, Ellen J. Weber, Christopher Barton, and Gary Young. 1998. Positive-Outcome Bias and Other Limitations in the Outcome of Research Abstracts Submitted to a Scientific Meeting. *JAMA* 280, 3 (July 1998), 254–257. https://doi.org/10.1001/jama.280.3.254

[20] Furui Cheng, Dongyu Liu, Fan Du, Yanna Lin, Alexandra Zytek, Haomin Li, Huamin Qu, and Kalyan Veeramachaneni. 2022. VBridge: Connecting the Dots Between Features and Data to Explain Healthcare Models. *IEEE Transactions on Visualization and Computer Graphics* 28, 1 (Jan. 2022), 378–388. https://doi.org/10.1109/TVCG.2021.3114836

[21] Furui Cheng, Yao Ming, and Huamin Qu. 2021. DECE: Decision Explorer with Counterfactual Explanations for Machine Learning Models. *IEEE Transactions on Visualization and Computer Graphics* 27, 2 (Feb. 2021), 1438–1447. https://doi.org/10.1109/TVCG.2020.3030342

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

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

[24] Dennis Collaris and Jarke J. van Wijk. 2020. ExplainExplore: Visual Exploration of Machine Learning Explanations. In *2020 IEEE Pacific Visualization Symposium (PacificVis)*. IEEE, Tianjin, China, 26–35. https://doi.org/10.1109/PacificVis48177.2020.7090 ISSN: 2165-8773.

[25] Jason A. Colquitt and Jessica B. Rodell. 2015. Measuring justice and fairness. In *The Oxford handbook of justice in the workplace*. Oxford University Press, New York, NY, US, 187–202. https://doi.org/10.1093/oxfordhb/9780199981410.013.8

[26] Richard Dazeley, Peter Vamplew, Cameron Foale, Charlotte Young, Sunil Aryal, and Francisco Cruz. 2021. Levels of explainable artificial intelligence for human-aligned conversational explanations. *Artificial Intelligence* 299 (Oct. 2021), 103525. https://doi.org/10.1016/j.artint.2021.103525

[27] Maartje M. A. de Graaf and Bertram F. Malle. 2017. How People Explain Action (and Autonomous Intelligent Systems Should Too). In *2017 AAAI Fall Symposia, Arlington, Virginia, USA, November 9-11, 2017*. AAAI Press, 19–26. https://aaai.org/ocs/index.php/FSS/FSS17/paper/view/16009

[28] John Dewey. 1903. Democracy in Education. *THE ELEMENTARY SCHOOL TEACHER* (1903), 12.

[29] Alan Dix and Geoffrey Ellis. 1998. Starting simple: adding value to static visualisation through simple interaction. In *Proceedings of the working conference on Advanced visual interfaces (AVI '98)*. Association for Computing Machinery, New York, NY, USA, 124–134. https://doi.org/10.1145/948496.948514

[30] Jonathan Dodge, Andrew A. Anderson, Matthew Olson, Rupika Dikkala, and Margaret Burnett. 2022. How Do People Rank Multiple Mutant Agents?. In *27th International Conference on Intelligent User Interfaces (IUI '22)*. Association for Computing Machinery, New York, NY, USA, 191–211. https://doi.org/10.1145/3490099.3511115

[31] Vicente Dominguez, Pablo Messina, Ivania Donoso-Guzmán, and Denis Parra. 2019. The effect of explanations and algorithmic accuracy on visual recommender systems of artistic images. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. ACM, Marina del Ray California, 408–416. https://doi.org/10.1145/3301275.3302274

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

[33] Filip Karlo Došilović, Mario Brčić, and Nikica Hlupić. 2018. Explainable artificial intelligence: A survey. In *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. IEEE,

Opatija, Croatia, 0210–0215. https://doi.org/10.23919/MIPRO.2018.8400040

[34] Chris Evans and Nicola J. Gibbons. 2007. The interactivity effect in multimedia learning. *Computers & Education* 49, 4 (Dec. 2007), 1147–1160. https://doi.org/10.1016/j.compedu.2006.01.008

[35] Shi Feng and Jordan Boyd-Graber. 2019. What can AI do for me? evaluating machine learning interpretations in cooperative play. In *Proceedings of the 24th International Conference on Intelligent User Interfaces (IUI '19)*. Association for Computing Machinery, New York, NY, USA, 229–239. https://doi.org/10.1145/3301275.3302265

[36] Juliana J. Ferreira and Mateus S. Monteiro. 2020. What Are People Doing About XAI User Experience? A Survey on AI Explainability Research and Practice. In *Design, User Experience, and Usability. Design for Contemporary Interactive Environments (Lecture Notes in Computer Science)*, Aaron Marcus and Elizabeth Rosenzweig (Eds.). Springer International Publishing, Cham, 56–73. https://doi.org/10.1007/978-3-030-49760-6_4

[37] James D. Foley, Foley Dan Van, Andries Van Dam, Steven K. Feiner, and John F. Hughes. 1996. *Computer Graphics: Principles and Practice*. Addison-Wesley Professional, USA.

[38] Bhavya Ghai, Q. Vera Liao, Yunfeng Zhang, Rachel Bellamy, and Klaus Mueller. 2021. Explainable Active Learning (XAL): Toward AI Explanations as Interfaces for Machine Teachers. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW3 (Jan. 2021), 1–28. https://doi.org/10.1145/3432934

[39] Azin Ghazimatin, Soumajit Pramanik, Rishiraj Saha Roy, and Gerhard Weikum. 2021. ELIXIR: Learning from User Feedback on Explanations to Improve Recommender Models. In *Proceedings of the Web Conference 2021 (WWW '21)*. Association for Computing Machinery, New York, NY, USA, 3850–3860. https://doi.org/10.1145/3442381.3449848

[40] H. P. Grice. 1975. *Logic and Conversation*. Brill. https://doi.org/10.1163/9789004368811_003 Pages: 41-58 Section: Speech Acts.

[41] Ziwei Gu, Jing Nathan Yan, and Jeffrey M. Rzeszotarski. 2021. Understanding User Sensemaking in Machine Learning Fairness Assessment Systems. In *Proceedings of the Web Conference 2021 (WWW '21)*. Association for Computing Machinery, New York, NY, USA, 658–668. https://doi.org/10.1145/3442381.3450092

[42] Lijie Guo, Elizabeth M. Daly, Oznur Alkan, Massimiliano Mattetti, Owen Cornec, and Bart Knijnenburg. 2022. Building Trust in Interactive Machine Learning via User Contributed Interpretable Rules. In *27th International Conference on Intelligent User Interfaces (IUI '22)*. Association for Computing Machinery, New York, NY, USA, 537–548. https://doi.org/10.1145/3490099.3511111

[43] Sam Hepenstal, Leishi Zhang, Neesha Kodagoda, and B. l. william Wong. 2021. Developing Conversational Agents for Use in Criminal Investigations. *ACM Transactions on Interactive Intelligent Systems* 11, 3-4 (Dec. 2021), 1–35. https://doi.org/10.1145/3444369

[44] Diana C. Hernandez-Bocanegra and Jürgen Ziegler. 2021. Conversational review-based explanations for recommender systems: Exploring users' query behavior. In *CUI 2021 - 3rd Conference on Conversational User Interfaces (CUI '21)*. Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3469595.3469596

[45] Germund Hesslow. 1988. The Problem of Causal Selection. In *Contemporary Science and Natural Explanation: Commonsense Conceptions of Causality*, Denis J. Hilton (Ed.). New York University Press.

[46] Robert R. Hoffman, Shane T. Mueller, Gary Klein, and Jordan Litman. 2019. Metrics for Explainable AI: Challenges and Prospects. http://arxiv.org/abs/1812.04608 arXiv:1812.04608 [cs].

[47] Fred Hohman, Andrew Head, Rich Caruana, Robert DeLine, and Steven M. Drucker. 2019. Gamut: A Design Probe to Understand How Data Scientists Understand Machine Learning Models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, Glasgow Scotland Uk, 1–13. https://doi.org/10.1145/3290605.3300809

[48] Maia Jacobs, Jeffrey He, Melanie F. Pradier, Barbara Lam, Andrew C. Ahn, Thomas H. McCoy, Roy H. Perlis, Finale Doshi-Velez, and Krzysztof Z. Gajos. 2021. Designing AI for Trust and Collaboration in Time-Constrained Medical Decisions: A Sociotechnical Lens. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–14. https://doi.org/10.1145/3411764.3445385

[49] Lars-Erik Janlert and Erik Stolterman. 2017. The Meaning of Interactivity—Some Proposals for Definitions and Measures. *Human–Computer Interaction* 32, 3 (May 2017), 103–138. https://doi.org/10.1080/07370024.2016.1226139 Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/07370024.2016.1226139.

[50] Shichao Jia, Zeyu Li, Nuo Chen, and Jiawan Zhang. 2022. Towards Visual Explainable Active Learning for Zero-Shot Classification. *IEEE Transactions on Visualization and Computer Graphics* 28, 1 (Jan. 2022), 791–801. https://doi.org/10.1109/TVCG.2021.3114793

[51] Zhuochen Jin, Shuyuan Cui, Shunan Guo, David Gotz, Jimeng Sun, and Nan Cao. 2020. CarePre: An Intelligent Clinical Decision Assistance System. *ACM Transactions on Computing for Healthcare* 1, 1 (March 2020), 6:1–6:20. https://doi.org/10.1145/3344258

[52] Daniel Kahneman and Amos Tversky. 1979. Prospect Theory: An Analysis of Decision under Risk. *Econometrica* 47, 2 (1979), 263–291. https://doi.org/10.2307/1914185 Publisher: [Wiley, Econometric Society].

[53] D.A. Keim. 2002. Information visualization and visual data mining. *IEEE Transactions on Visualization and Computer Graphics* 8, 1 (Jan. 2002), 1–8. https://doi.org/10.1109/2945.981847

[54] Carmel Kent, Esther Laslo, and Sheizaf Rafaeli. 2016. Interactivity in online discussions and learning outcomes. *Computers & Education* 97 (June 2016), 116–128. https://doi.org/10.1016/j.compedu.2016.03.002

[55] Anjali Khurana, Parsa Alamzadeh, and Parmit K. Chilana. 2021. ChatrEx: Designing Explainable Chatbot Interfaces for Enhancing Usefulness, Transparency, and Trust. In *2021 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*. IEEE, St Louis, MO, USA, 1–11. https://doi.org/10.1109/VL/HCC51201.2021.9576440 ISSN: 1943-6106.

[56] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. 2018. Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). https://doi.org/10.48550/arXiv.1711.11279 arXiv:1711.11279 [stat].

[57] Chris Kim, Xiao Lin, Christopher Collins, Graham W. Taylor, and Mohamed R. Amer. 2021. Learn, Generate, Rank, Explain: A Case Study of Visual Explanation by Generative Machine Learning. *ACM Transactions on Interactive Intelligent Systems* 11, 3-4 (Sept. 2021), 23:1–23:34. https://doi.org/10.1145/3465407

[58] Bart P. Knijnenburg, Martijn C. Willemsen, Zeno Gantner, Hakan Soncu, and Chris Newell. 2012. Explaining the user experience of recommender systems. *User Modeling and User-Adapted Interaction* 22, 4 (Oct. 2012), 441–504. https://doi.org/10.1007/s11257-011-9118-4

[59] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. 2020. Concept Bottleneck Models. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020 (Proceedings of Machine Learning Research, Vol. 119)*. PMLR, Virtual Event, 5338–5348. http://proceedings.mlr.press/v119/koh20a.html

[60] Yubo Kou and Xinning Gui. 2020. Mediating Community-AI Interaction through Situated Explanation: The Case of AI-Led Moderation. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (Oct. 2020), 102:1–102:27. https://doi.org/10.1145/3415173

[61] Pigi Kouki, James Schaffer, Jay Pujara, John O'Donovan, and Lise Getoor. 2019. Personalized explanations for hybrid recommender systems. In *Proceedings of the 24th International Conference on Intelligent User Interfaces (IUI '19)*. Association for Computing Machinery, New York, NY, USA, 379–390. https://doi.org/10.1145/3301275.3302306

[62] Maria Kouvela, Ilias Dimitriadis, and Athena Vakali. 2020. Bot-Detective: An explainable Twitter bot detection service with crowdsourcing functionalities. In *Proceedings of the 12th International Conference on Management of Digital EcoSystems (MEDES '20)*. Association for Computing Machinery, New York, NY, USA, 55–63. https://doi.org/10.1145/3415958.3433075

[63] Josua Krause, Adam Perer, and Kenney Ng. 2016. Interacting with Predictions: Visual Inspection of Black-box Machine Learning Models. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. Association for Computing Machinery, New York, NY, USA, 5686–5697. https://doi.org/10.1145/2858036.2858529

[64] Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. 2015. Principles of Explanatory Debugging to Personalize Interactive Machine Learning. In *Proceedings of the 20th International Conference on Intelligent User Interfaces (IUI '15)*. Association for Computing Machinery, New York, NY, USA, 126–137. https://doi.org/10.1145/2678025.2701399

[65] Bum Chul Kwon, Min-Je Choi, Joanne Taery Kim, Edward Choi, Young Bin Kim, Soonwook Kwon, Jimeng Sun, and Jaegul Choo. 2019. RetainVis: Visual Analytics with Interpretable and Interactive Recurrent Neural Networks on Electronic Medical Records. *IEEE Transactions on Visualization and Computer Graphics* 25, 1 (Jan. 2019), 299–309. https://doi.org/10.1109/TVCG.2018.2865027

[66] Vivian Lai, Chacha Chen, Q. Vera Liao, Alison Smith-Renner, and Chenhao Tan. 2021. Towards a Science of Human-AI Decision Making: A Survey of Empirical Studies. https://doi.org/10.48550/arXiv.2112.11471 arXiv:2112.11471 [cs].

[67] Markus Langer, Daniel Oster, Timo Speith, Holger Hermanns, Lena Kästner, Eva Schmidt, Andreas Sesing, and Kevin Baum. 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 (July 2021), 103473. https://doi.org/10.1016/j.artint.2021.103473

[68] Min Kyung Lee, Anuraag Jain, Hea Jin Cha, Shashank Ojha, and Daniel Kusbit. 2019. Procedural Justice in Algorithmic Fairness: Leveraging Transparency and Outcome Control for Fair Algorithmic Mediation. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (Nov. 2019), 182:1–182:26. https://doi.org/10.1145/3359284

[69] James R. Lewis. 1991. Psychometric evaluation of an after-scenario questionnaire for computer usability studies: the ASQ. *ACM SIGCHI Bulletin* 23, 1 (Jan. 1991), 78–81. https://doi.org/10.1145/122672.122692

[70] Q. Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–15.

https://doi.org/10.1145/3313831.3376590

[71] Brian Y. Lim and Anind K. Dey. 2009. Assessing demand for intelligibility in context-aware applications. In *Proceedings of the 11th international conference on Ubiquitous computing (UbiComp '09)*. Association for Computing Machinery, New York, NY, USA, 195–204. https://doi.org/10.1145/1620545.1620576

[72] Peter Lipton. 1990. Contrastive Explanation*. *Royal Institute of Philosophy Supplements* 27 (March 1990), 247–266. https://doi.org/10.1017/S1358246100005130 Publisher: Cambridge University Press.

[73] Han Liu, Vivian Lai, and Chenhao Tan. 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 (Oct. 2021), 408:1–408:45. https://doi.org/10.1145/3479552

[74] D. Mcknight, Michelle Carter, Jason Thatcher, and Paul Clay. 2011. Trust in a specific technology: An Investigation of its Components and Measures. *ACM Transactions on Management Information Systems* 2 (June 2011), 12–32. https://doi.org/10.1145/1985347.1985353

[75] Gaspar Isaac Melsión, Ilaria Torre, Eva Vidal, and Iolanda Leite. 2021. Using Explainability to Help Children UnderstandGender Bias in AI. In *Interaction Design and Children*. ACM, Athens Greece, 87–99. https://doi.org/10.1145/3459990.3460719

[76] Martijn Millecamp, Nyi Nyi Htun, Cristina Conati, and Katrien Verbert. 2019. To explain or not to explain: the effects of personal characteristics when explaining music recommendations. In *Proceedings of the 24th International Conference on Intelligent User Interfaces (IUI '19)*. Association for Computing Machinery, New York, NY, USA, 397–407. https://doi.org/10.1145/3301275.3302313

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

[78] Yao Ming, Huamin Qu, and Enrico Bertini. 2019. RuleMatrix: Visualizing and Understanding Classifiers with Rules. *IEEE Transactions on Visualization and Computer Graphics* 25, 1 (Jan. 2019), 342–352. https://doi.org/10.1109/TVCG.2018.2864812

[79] Akira Miyake and Priti Shah (Eds.). 1999. *Models of working memory: Mechanisms of active maintenance and executive control*. Cambridge University Press, New York, NY, US. https://doi.org/10.1017/CBO9781139174909 Pages: xx, 506.

[80] David Moher, Alessandro Liberati, Jennifer Tetzlaff, Douglas G. Altman, and The PRISMA Group. 2009. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *PLoS Medicine* 6, 7 (July 2009), e1000097. https://doi.org/10.1371/journal.pmed.1000097

[81] Cecily Morrison, Kit Huckvale, Bob Corish, Richard Banks, Martin Grayson, Jonas Dorn, Abigail Sellen, and Sân Lindley. 2018. Visualizing Ubiquitously Sensed Measures of Motor Ability in Multiple Sclerosis: Reflections on Communicating Machine Learning in Practice. *ACM Transactions on Interactive Intelligent Systems* 8, 2 (July 2018), 1–28. https://doi.org/10.1145/3181670

[82] C. D. Mulrow. 1994. Systematic Reviews: Rationale for systematic reviews. *BMJ* 309, 6954 (Sept. 1994), 597–599. https://doi.org/10.1136/bmj.309.6954.597 Publisher: British Medical Journal Publishing Group Section: Education and debate.

[83] Zachary Munn, Micah D. J. Peters, Cindy Stern, Catalin Tufanaru, Alexa McArthur, and Edoardo Aromataris. 2018. Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. *BMC Medical Research Methodology* 18, 1 (Nov. 2018), 143. https://doi.org/10.1186/s12874-018-0611-x

[84] Mohammad Naiseh, Reem S. Al-Mansoori, Dena Al-Thani, Nan Jiang, and Raian Ali. 2021. Nudging through Friction: An Approach for Calibrating Trust in Explainable AI. In *2021 8th International Conference on Behavioral and Social Computing (BESC)*. IEEE, Doha, Qatar, 1–5. https://doi.org/10.1109/BESC53957.2021.9635271

[85] Mahsan Nourani, Chiradeep Roy, Jeremy E Block, Donald R Honeycutt, Tahrima Rahman, Eric Ragan, and Vibhav Gogate. 2021. Anchoring Bias Affects Mental Model Formation and User Reliance in Explainable AI Systems. In *26th International Conference on Intelligent User Interfaces*. Association for Computing Machinery, New York, NY, USA, 340–350. https://doi.org/10.1145/3397481.3450639

[86] Heather L. O'Brien, Paul Cairns, and Mark Hall. 2018. A practical approach to measuring user engagement with the refined user engagement scale (UES) and new UES short form. *International Journal of Human-Computer Studies* 112 (April 2018), 28–39. https://doi.org/10.1016/j.ijhcs.2018.01.004

[87] Matthew J Page, Joanne E McKenzie, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M Tetzlaff, Elie A Akl, Sue E Brennan, et al. 2021. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *Systematic reviews* 10, 1 (2021), 1–11.

[88] Sean Penney, Jonathan Dodge, Claudia Hilderbrand, Andrew Anderson, Logan Simpson, and Margaret Burnett. 2018. Toward Foraging for Understanding of StarCraft Agents: An Empirical Study. In *23rd International Conference on Intelligent User Interfaces (IUI '18)*. Association for Computing Machinery, New York, NY, USA, 225–237. https://doi.org/10.1145/3172944.3172946

[89] Sayantan Polley, Suhita Ghosh, Marcus Thiel, Michael Kotzyba, and Andreas Nürnberger. 2020. SIMFIC: An Explainable Book Search Companion. In *2020 IEEE International Conference on Human-Machine Systems (ICHMS)*. IEEE, Rome, Italy, 1–6. https://doi.org/10.1109/ICHMS49158.2020.9209581

[90] Gabriëlle Ras, Marcel van Gerven, and Pim Haselager. 2018. Explanation Methods in Deep Learning: Users, Values, Concerns and Challenges. In *Explainable and Interpretable Models in Computer Vision and Machine Learning*, Hugo Jair Escalante, Sergio Escalera, Isabelle Guyon, Xavier Baró, Yağmur Güçlütürk, Umut Güçlü, and Marcel van Gerven (Eds.). Springer International Publishing, Cham, 19–36. https://doi.org/10.1007/978-3-319-98131-4_2

[91] Juan Rebanal, Jordan Combitsis, Yuqi Tang, and Xiang 'Anthony' Chen. 2021. XAlgo: a Design Probe of Explaining Algorithms' Internal States via Question-Answering. In *26th International Conference on Intelligent User Interfaces (IUI '21)*. Association for Computing Machinery, New York, NY, USA, 329–339. https://doi.org/10.1145/3397481.3450676

[92] Dent M. Rhodes and Janet White Azbell. 1985. Designing Interactive Video Instruction Professionally. *Training and Development Journal* 39, 12 (1985), 31–33.

[93] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*. Association for Computing Machinery, New York, NY, USA, 1135–1144. https://doi.org/10.1145/2939672.2939778

[94] Andrew Ross, Nina Chen, Elisa Zhao Hang, Elena L. Glassman, and Finale Doshi-Velez. 2021. Evaluating the Interpretability of Generative Models by Interactive Reconstruction. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–15. https://doi.org/10.1145/3411764.3445296

[95] Steven F. Roth and Joe Mattis. 1990. Data characterization for intelligent graphics presentation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '90)*. Association for Computing Machinery, New York, NY, USA, 193–200. https://doi.org/10.1145/97243.97273

[96] Denise M. Rousseau, Sim B. Sitkin, Ronald S. Burt, and Colin Camerer. 1998. Introduction to Special Topic Forum: Not so Different after All: A Cross-Discipline View of Trust. *The Academy of Management Review* 23, 3 (1998), 393–404. http://www.jstor.org/stable/259285

[97] Maria Roussou. 2004. Learning by doing and learning through play: an exploration of interactivity in virtual environments for children. *Computers in Entertainment* 2, 1 (Jan. 2004), 10. https://doi.org/10.1145/973801.973818

[98] James Schaffer, Prasanna Giridhar, Debra Jones, Tobias Höllerer, Tarek Abdelzaher, and John O'Donovan. 2015. Getting the Message? A Study of Explanation Interfaces for Microblog Data Analysis. In *Proceedings of the 20th International Conference on Intelligent User Interfaces (IUI '15)*. Association for Computing Machinery, New York, NY, USA, 345–356. https://doi.org/10.1145/2678025.2701406

[99] Kara Schick-Makaroff, Marjorie MacDonald, Marilyn Plummer, Judy Burgess, and Wendy Neander. 2016. What Synthesis Methodology Should I Use? A Review and Analysis of Approaches to Research Synthesis. *AIMS public health* 3, 1 (March 2016), 172–215. https://doi.org/10.3934/publichealth.2016.1.172

[100] Johanes Schneider and Joshua Handali. 2019. Personalized explanation in machine learning: A conceptualization. In *Proceedings of the European Conference on Information Systems (ECIS 2019)*. arXiv, Stockholm-Uppsala, Sweden. https://doi.org/10.48550/arXiv.1901.00770 arXiv:1901.00770 [cs, stat].

[101] Richard Schwier and Earl R. Misanchuk. 1993. *Interactive Multimedia Instruction*. Educational Technology.

[102] Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. 2019. What makes a good conversation? How controllable attributes affect human judgments. https://doi.org/10.48550/arXiv.1902.08654 arXiv:1902.08654 [cs].

[103] Rita Sevastjanova, Wolfgang Jentner, Fabian Sperrle, Rebecca Kehlbeck, Jürgen Bernard, and Mennatallah El-assady. 2021. QuestionComb: A Gamification Approach for the Visual Explanation of Linguistic Phenomena through Interactive Labeling. *ACM Transactions on Interactive Intelligent Systems* 11, 3-4 (Dec. 2021), 1–38. https://doi.org/10.1145/3429448

[104] Lei Shi, Zhiyang Teng, Le Wang, Yue Zhang, and Alexander Binder. 2019. DeepClue: Visual Interpretation of Text-Based Deep Stock Prediction. *IEEE Transactions on Knowledge and Data Engineering* 31, 6 (June 2019), 1094–1108. https://doi.org/10.1109/TKDE.2018.2854193

[105] Rod Sims. 1997. Interactivity: A forgotten art? *Computers in Human Behavior* 13, 2 (May 1997), 157–180. https://doi.org/10.1016/S0747-5632(97)00004-6

[106] Dylan Slack, Satyapriya Krishna, Himabindu Lakkaraju, and Sameer Singh. 2022. TalkToModel: Explaining Machine Learning Models with Interactive Natural Language Conversations. https://doi.org/10.48550/arXiv.2207.04154 arXiv:2207.04154 [cs].

[107] Kacper Sokol and Peter Flach. 2020. Explainability fact sheets: a framework for systematic assessment of explainable approaches. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM, Barcelona Spain, 56–67. https://doi.org/10.1145/3351095.3372870

[108] Kacper Sokol and Peter Flach. 2020. One Explanation Does Not Fit All: The Promise of Interactive Explanations for Machine Learning Transparency. *KI - Künstliche Intelligenz* 34, 2 (June 2020), 235–250. https://doi.org/10.1007/s13218-

020-00637-y arXiv:2001.09734 [cs, stat].

[109] Francesco Sovrano and Fabio Vitali. 2021. From Philosophy to Interfaces: an Explanatory Method and a Tool Inspired by Achinstein's Theory of Explanation. In *26th International Conference on Intelligent User Interfaces*. ACM, College Station TX USA, 81–91. https://doi.org/10.1145/3397481.3450655

[110] Thilo Spinner, Udo Schlegel, Hanna Schäfer, and Mennatallah El-Assady. 2020. explAIner: A Visual Analytics Framework for Interactive and Explainable Machine Learning. *IEEE Transactions on Visualization and Computer Graphics* 26, 1 (Jan. 2020), 1064–1074. https://doi.org/10.1109/TVCG.2019.2934629

[111] Aaron Springer and Steve Whittaker. 2019. Progressive disclosure: empirically motivated approaches to designing effective transparency. In *Proceedings of the 24th International Conference on Intelligent User Interfaces (IUI '19)*. Association for Computing Machinery, New York, NY, USA, 107–120. https://doi.org/10.1145/3301275.3302322

[112] Jonathan Steuer. 1992. Defining Virtual Reality: Dimensions Determining Telepresence. *Journal of Communication* (1992), 73–93.

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

[114] S. Shyam Sundar, Qian Xu, and Saraswathi Bellur. 2010. Designing interactivity in media interfaces: a communications perspective. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10)*. Association for Computing Machinery, New York, NY, USA, 2247–2256. https://doi.org/10.1145/1753326.1753666

[115] Harini Suresh, Kathleen M Lewis, John Guttag, and Arvind Satyanarayan. 2022. Intuitively Assessing ML Model Reliability through Example-Based Explanations and Editing Model Inputs. In *27th International Conference on Intelligent User Interfaces (IUI '22)*. Association for Computing Machinery, New York, NY, USA, 767–781. https://doi.org/10.1145/3490099.3511160

[116] Maxwell Szymanski, Martijn Millecamp, and Katrien Verbert. 2021. Visual, textual or hybrid: the effect of user expertise on different explanations. In *26th International Conference on Intelligent User Interfaces*. ACM, College Station TX USA, 109–119. https://doi.org/10.1145/3397481.3450662

[117] Nava Tintarev. 2007. Explanations of recommendations. In *Proceedings of the 2007 ACM conference on Recommender systems (RecSys '07)*. Association for Computing Machinery, New York, NY, USA, 203–206. https://doi.org/10.1145/1297231.1297275

[118] Andrea C. Tricco, Erin Lillie, Wasifa Zarin, Kelly K. O'Brien, Heather Colquhoun, Danielle Levac, David Moher, Micah D.J. Peters, Tanya Horsley, Laura Weeks, Susanne Hempel, Elie A. Akl, Christine Chang, Jessie McGowan, Lesley Stewart, Lisa Hartling, Adrian Aldcroft, Michael G. Wilson, Chantelle Garritty, Simon Lewin, Christina M. Godfrey, Marilyn T. Macdonald, Etienne V. Langlois, Karla Soares-Weiser, Jo Moriarty, Tammy Clifford, Özge Tunçalp, and Sharon E. Straus. 2018. PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. *Annals of Internal Medicine* 169, 7 (Oct. 2018), 467–473. https://doi.org/10.7326/M18-0850 Publisher: American College of Physicians.

[119] Chun-Hua Tsai, Yue You, Xinning Gui, Yubo Kou, and John M. Carroll. 2021. Exploring and Promoting Diagnostic Transparency and Explainability in Online Symptom Checkers. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–17. https://doi.org/10.1145/3411764.3445101

[120] Betty Vandenbosch and Michael J. Ginzberg. 1996. Lotus Notes® and Collaboration: Plus ça change... *Journal of Management Information Systems* 13, 3 (Dec. 1996), 65–81. https://doi.org/10.1080/07421222.1996.11518134 Publisher: Routledge _eprint: https://doi.org/10.1080/07421222.1996.11518134.

[121] Marco Virgolin, Andrea De Lorenzo, Francesca Randone, Eric Medvet, and Mattias Wahde. 2021. Model learning with personalized interpretability estimation (ML-PIE). In *Proceedings of the Genetic and Evolutionary Computation Conference Companion (GECCO '21)*. Association for Computing Machinery, New York, NY, USA, 1355–1364. https://doi.org/10.1145/3449726.3463166

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

[123] Jane Webster and Richard T. Watson. 2002. Analyzing the Past to Prepare for the Future: Writing a Literature Review. *MIS Quarterly* 26, 2 (2002), xiii–xxiii. https://www.jstor.org/stable/4132319 Publisher: Management Information Systems Research Center, University of Minnesota.

[124] Daricia Wilkinson, Öznur Alkan, Q. Vera Liao, Massimiliano Mattetti, Inge Vejsbjerg, Bart P. Knijnenburg, and Elizabeth Daly. 2021. Why or Why Not? The Effect of Justification Styles on Chatbot Recommendations. *ACM Transactions on Information Systems* 39, 4 (Oct. 2021), 1–21. https://doi.org/10.1145/3441715

[125] Leland Wilkinson. 2005. Introduction. In *The Grammar of Graphics*. Springer, New York, NY, 1–19. https://doi.org/10.1007/0-387-28695-0_1

[126] Tongshuang Wu, Marco Túlio Ribeiro, Jeffrey Heer, and Daniel S. Weld. 2021. Polyjuice: Generating Counterfactuals for Explaining, Evaluating, and Improving Models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP*, Chengqing Zong, Fei Xia, Wenjie Li 0002, and Roberto Navigli (Eds.). Association for Computational Linguistics, Virtual Event, 6707–6723. https://aclanthology.org/2021.acl-long.523

[127] Jing Nathan Yan, Ziwei Gu, Hubert Lin, and Jeffrey M. Rzeszotarski. 2020. Silva: Interactively Assessing Machine Learning Fairness Using Causality. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376447

[128] Ji Soo Yi, Youn ah Kang, John Stasko, and J.A. Jacko. 2007. Toward a Deeper Understanding of the Role of Interaction in Information Visualization. *IEEE Transactions on Visualization and Computer Graphics* 13, 6 (Nov. 2007), 1224–1231. https://doi.org/10.1109/TVCG.2007.70515

APPENDIX

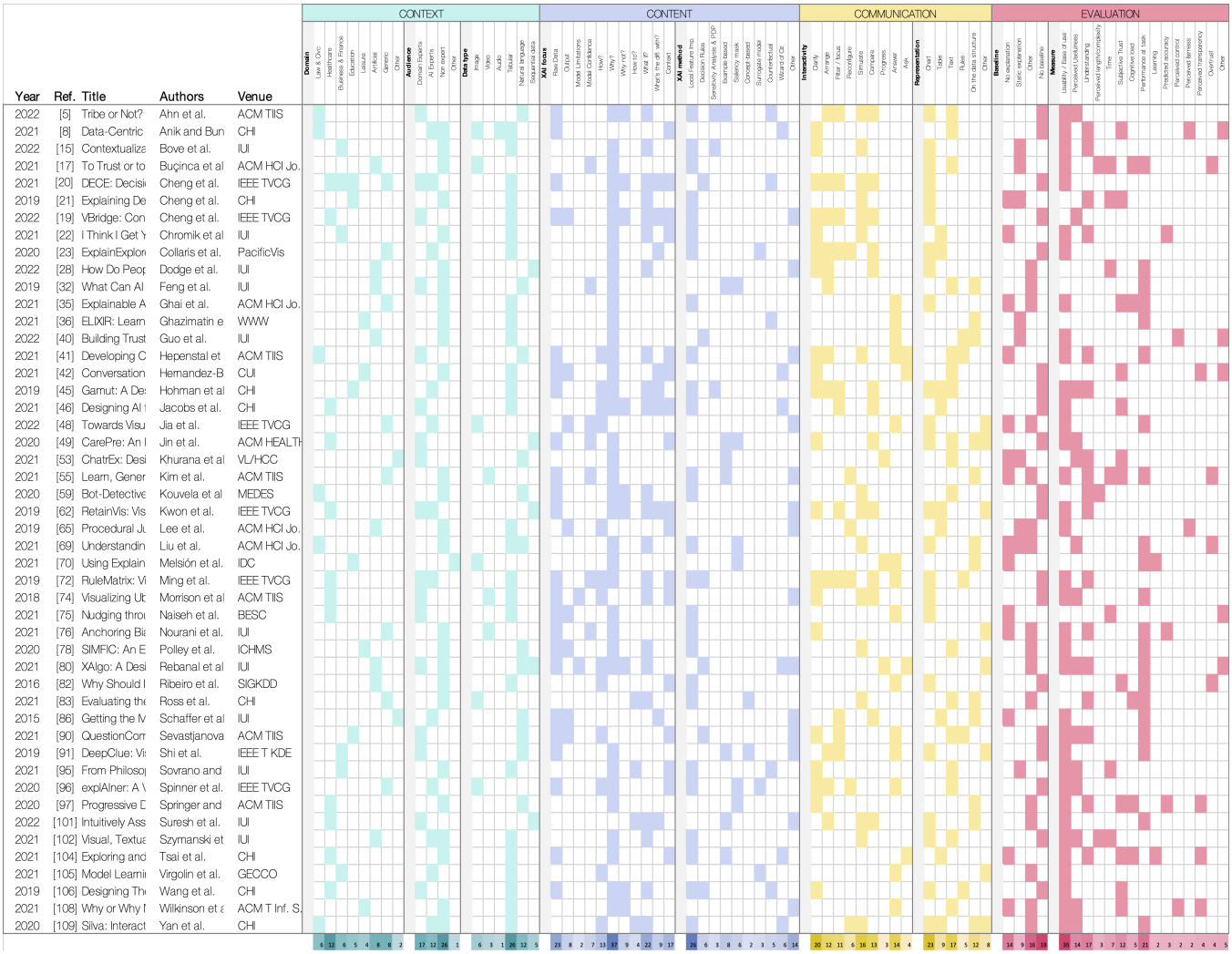


Figure 4: The concept matrix [123] resulting from our survey of 48 empirical studies evaluating interactive explainability systems. Our analysis focused on four dimensions: the context in which the explanations were found, their content, how they were communicated to users, including the interaction technique used, and the effects they had on users. The last row at the bottom of the matrix shows the total number of items for each sub-dimension. The design of this concept matrix was inspired from [10].

Table 4: Example references of the interactivity categories with corresponding description of the interactive feature.

Level	Category	Reference	Description of the interactive feature
Select	*Clarify*	Fig. 3, Sovrano and Vitali *et al.* [109]	Users can click on the concepts present in an explanation to see definitions. Here, users could click on "Months since most recent credit inquiry not within the last 7 days" to see a pop-up window appear with the definition of the concept. The definition itself includes links towards definitions of other terms, like "inquiry".
		Fig. 1, Anik and Bunt [7]	Users see a menu with categories about the AI system "[Data] Collection", "Demographics", "Recommended Usage", Potential Issues" and "General Information" and can explore these categories at their pace to learn more about them.
	Arrange	Fig. 1B "Instance view", DECE by Cheng *et al.* [21]	Users can choose the number of counterfactuals, the number of features used in explanations and the type of constraints used to generate counterfactuals.
		Fig. 1, top-left, GAMUT by Hohman *et al.* [47]	Users can choose to normalize the axes of the graphs presented as explanations, hide all histograms in the explanation interface, or hide the dashed zeroline appearing in the presented graphs.
	Filter/focus	Fig. 2A "Feature Sidebar", GAMUT by Hohman *et al.* [47]	Users can sort or choose to hide or show the different input features appearing in feature importance explanations.
		Fig. 1, "Controls", RULEMATRIX by Ming *et al.* [78]	Users can filter rule-based explanations by the level of minimal evidence they present or by the level of fidelity to the original model.
Mutate	*Reconfigure*	Fig. 1, "Controls", RULEMATRIX by Ming *et al.* [78]	Users can choose the dataset used: "train, test, sample test or sample train"
		Fig. 4A "Configuration view", ExplainExplore by Collaris and van Wijk [24]	Users can choose the dataset and AI model used through drop-down lists in a control pane on the left of the interface.
	Compare	Fig. 1, TribalGram by Ahn *et al.* [4]	In the different explanations presented in the interface, users can see and compare the "blue camp" and the "red camp"' data groups, each represented respectively in blue or red.'
		Fig. 2C "Instance explanation", GAMUT by Hohman *et al.* [47]	The interface presents two waterfall charts stacked on top of each other, each explaining the price prediction for two different houses. The waterfall charts show the importance of various features on the prediction—most important features on the right. The stacked charts share the same x-axis with those features. The user can then easily compare the feature importance for each house instance.
	Simulate	Fig. 3, Ross *et al.* [94]	Users can vary the value of six input concepts that parametrize the shape of an output image representing a handwritten number using sliders or radio buttons, and see the effect on the output image in real time.
		Fig. 1a, Cheng *et al.* [22]	Users can vary the value of the inputs of a school admission decision-making algorithm. Quantitative inputs can be changed through sliders (GRE scores, GPA...) and Qualitative inputs through drop-downs (weak, medium or strong letter of recommendation).
Dialogue with	*Progress*	Fig. 4a, Melsión *et al.* [75]	Users can browse through example images. The most "important" image portions for the algorithm's prediction (here the gender of the person in the image) are highlighted in red.
		Fig. 1, ChatrEx by Khurana *et al.* [55]	Users can navigate through a three-step explanation about the presented system using next and previous buttons.
	Answer	Fig. 4b, Melsión *et al.* [75]	Users can select what they think is the most important part of an image by clicking directly on circled areas of the image (in this case, either a surfer, a surfboard or the beach).
		Fig. 3, Guo *et al.* [42]	Users are asked to enter their prediction about the outcome of a Tic-Tac-Toe game board (either click on "X wins" or "O wins"). Users are also asked how they would like to change the proposed rule for the AI's predicted outcome, by picking a cell on the board game and rule operations (!= or =).
	Ask	Fig. 4, bottom right, Hepenstal *et al.* [43]	Users can ask the criminal investigation system questions like "what do you know about susan leech", and the system answers with suggestions like "IDSusan Leech (Person)" or "Susan Leech (Information Object)".
		Fig. 3, Hernandez-Bocanegra and Ziegler [44]	Users can chat with a "recommendation explainer" chatbot that explains hotel recommendations, ask questions like "why is hotel julian in a good location" and get answers in natural language.

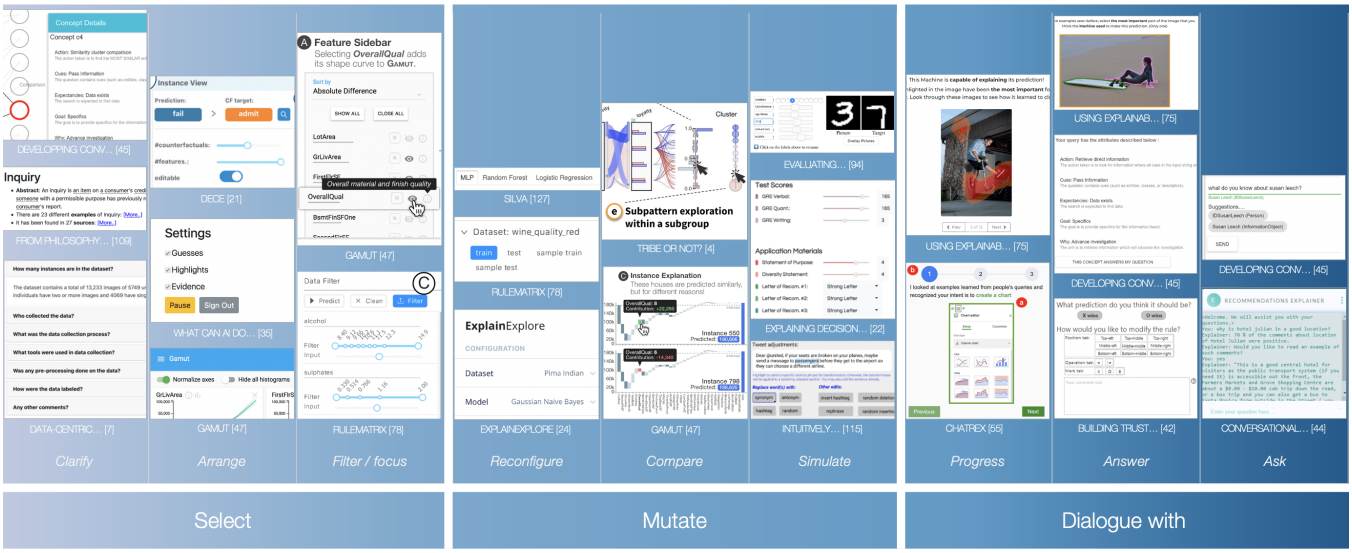


Figure 5: Examples of the interactivity categories through selected screenshots from the corpus. All images are copyrighted to their authors.