How Cognitive Biases Affect XAI-assisted Decision-making: A Systematic Review

Astrid Bertrand, Rafik Belloum, James Eagan, Winston Maxwell

How Cognitive Biases Affect XAI-assisted Decision-making: A Systematic Review

Astrid Bertrand
i3, Télécom Paris, Institut Polytechnique de Paris, Palaiseau

Rafik Belloum
LTCI, Télécom Paris, Institut Polytechnique de Paris
Palaiseau

James R. Eagan
LTCI, Télécom Paris, Institut Polytechnique de Paris
Palaiseau

Winston Maxwell
i3, Télécom Paris, Institut Polytechnique de Paris, Palaiseau

ABSTRACT

The field of eXplainable Artificial Intelligence (XAI) aims to bring transparency to complex AI systems. Although it is usually considered an essentially technical field, effort has been made recently to better understand users' human explanation methods and cognitive constraints. Despite these advances, the community lacks a general vision of what and how cognitive biases affect explainability systems. To address this gap, we present a heuristic map which matches human cognitive biases with explainability techniques from the XAI literature, structured around XAI-aided decision-making. We identify four main ways cognitive biases affect or are affected by XAI systems: 1) cognitive biases affect how XAI methods are designed, 2) they can distort how XAI techniques are evaluated in user studies, 3) some cognitive biases can be successfully mitigated by XAI techniques, and, on the contrary, 4) some cognitive biases can be exacerbated by XAI techniques. We construct this heuristic map through the systematic review of 37 papers—drawn from a corpus of 285—that reveal cognitive biases in XAI systems, including the explainability method and the user and task types in which they arise. We use the findings from our review to structure directions for future XAI systems to better align with people's cognitive processes.

CCS CONCEPTS

• Human-centered computing ~ Human computer interaction (HCI) ~ HCI theory, concepts and models • Computing methodologies ~ Artificial intelligence ~ Cognitive science

KEYWORDS

Explainability, explainable AI, cognitive bias, human-centered AI, XAI.

AIES '22, August 1–3, 2022, Oxford, United Kingdom.

ACM ISBN 978-1-4503-9247-1/22/08...$15.00.
https://doi.org/10.1145/3514094.3534164

ACM Reference format:

Astrid Bertrand, Rafik Belloum, James R. Eagan and Winston Maxwell. 2022. How Cognitive Biases Affect XAI-assisted Decision-making: A Systematic Review. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (AIES'22), August 1–3, 2022, Oxford, United Kingdom.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3514094.3534164

1 Introduction

In recent years, XAI has made significant breakthroughs in making opaque models more transparent [1]–[4]. Although many studies have shown that XAI methods can improve users' understanding of black-box models [5]–[7], recent empirical studies have drawn attention to obstacles resulting from a mismatch between people's cognitive constraints and current XAI techniques. For example, explanations can lead to unjustified trust in AI recommendations. Eiband *et al.* [8], show that placebic explanations elicit a similar level of trust as real explanations. Other work [9]–[11] shows that explanations can cause reasoning errors such as backward reasoning and confirmation bias [12]. These results highlight the danger of deploying AI system explanations in high-stakes settings without ensuring that they align with cognitive processes of users.

To address this gap, researchers [12]–[14] have been using insights about people's cognition and behaviors such as dual-process theory [15], which states that people rely on heuristics and cognitive biases to process information and make decisions. In the 1980s, Amos Tversky and Daniel Kahneman [16] defined cognitive biases as "systematic error in judgment and decision-making common to all human beings which can be due to cognitive limitations, motivational factors, and/or adaptations to natural environments." Leveraging Kahneman's dual process theory [1], Kliegr et al. [17] reviewed the effects of cognitive biases on the interpretation of AI models and provide a rich analysis of over 20 different biases. That work, however, focuses on rule-based explanations. Broader coverage of the cognitive biases related to XAI techniques in general is needed. In turn, Wang et al. [12] propose operational pathways between users' reasoning needs and XAI methodologies. They describe how people reason when explaining and review some common cognitive biases along with ways they can be mitigated. This is one of the most actionable contributions to date to link human

reasonings to XAI solutions. However, this work does not comprehensively cover the cognitive biases that may arise in the presence of XAI. While there are growing efforts from researchers [18]–[20] to tie cognitive science literature to a mostly technical XAI field, more research is needed to identify what kind of cognitive bias and heuristics are involved in the explanation process, and whether and how to leverage people's heuristics to improve XAI systems.

In this paper, we analyze how the field of XAI has been dealing with human cognitive biases and constraints, and we provide a research agenda that summarizes promising mitigation strategies and research directions to support human critical thinking. To this end, we conducted a systematic review of 37 papers, guided by the following five research questions: **RQ1:** What cognitive biases have been studied in the XAI literature? **RQ2:** In which contexts (*e.g.*, explainability method, human expertise, tasks type) do these cognitive biases arise? **RQ3:** How to mitigate negative biases and leverage appropriate biases to improve XAI systems? **RQ4:** What evaluation methods have been used to detect cognitive biases (specific to each bias)? **RQ5:** What are the stated future research directions and challenges identified by the scientific community?

This systematic review contributes the following to the XAI community:

- An identification of 53 cognitive biases mentioned so far in the XAI literature through a systematic methodology, providing an overview of the context in which these biases occur: with which XAI technique (e.g., counterfactual explanations), user type (domain expert, AI expert or lay users) and AI-assisted task (e.g., medical diagnosis).
- A heuristic map based on a systematic analysis of these cognitive biases, revealing four main ways they affect or are affected by XAI systems: 1) cognitive biases affect how XAI methods are designed, 2) they can distort how XAI techniques are evaluated in user studies, 3) some cognitive biases can be successfully mitigated by XAI techniques, and, on the contrary, 4) some cognitive biases can be exacerbated by XAI techniques.
- A research agenda for the XAI community to consider people's cognitive needs, addressing the key concerns and challenges we ran into during our review (e.g., improve perception of the user's reactions to XAI).

To the best of our knowledge, there is not yet a comprehensive review of how cognitive biases have been accounted for so far in the XAI literature. A systematic analysis like the one we present appears necessary to summarize findings on how cognitive biases interfere with explanations, how to address them, and to highlight promising directions concerning the integration of cognitive processes in XAI systems.

In this work, we consider cognitive biases not only in terms of "errors" (*e.g.*, automation bias that leads to inappropriate trust in AI modes) but also as the cognitive constraints that are inherent in the human explanation process (*e.g.*, homunculus bias, according to which people tend to attribute human traits to systems, and therefore expect explanations in the form of dialogues).

This paper is structured in four sections. Section 2 details the method used for the systematic literature review. Section 3 presents the results from the systematic review based on the proposed methodology and the identified research questions. Section 4 highlights the discussion of findings from the corpus and proposes several implications for future work. Section 5 concludes this work.

2 Methodology

In this section, we detail the method used for the systematic literature review and how we selected the papers for inclusion.

Our aim was to give a sense of how the XAI literature has addressed the notion of cognitive biases so far. We therefore relied on a keyword-based approach, which essentially has the advantage of ensuring transparency, reproducibility and, also, leading to more comprehensive results by sampling a wide range of work. However, it is possible that some XAI articles have addressed the notion of cognitive biases in different terms, referring to specific types of cognitive bias; we could not include all possible types of cognitive biases as keywords, since there are over 200. We also did not want to focus the investigation on specific types of bias in order to provide a more representative view of the different cognitive biases discussed in XAI. In addition, because we conducted our searches on ACM, IEEE, and Scopus, we may have missed other relevant work from other sources. To address these limitations, we supplemented the keyword-search with selected papers addressing cognitive biases in XAI drawn from two authors' knowledge of the XAI field.

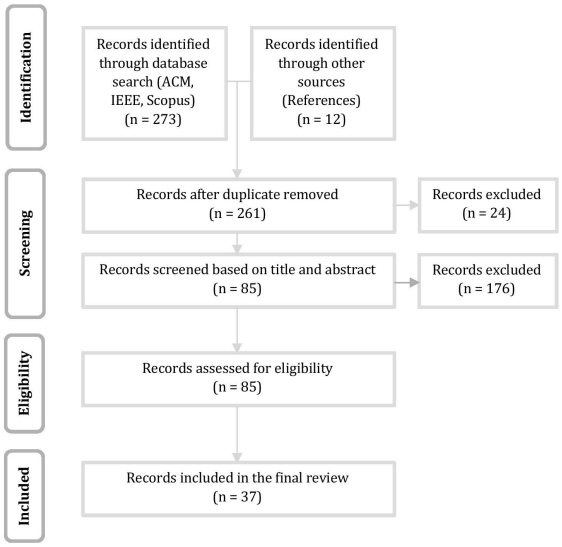

Figure 1: PRISMA flow diagram [21] on how the final corpus was curated (n = 37).

To guide the development of our systematic literature review we used the reporting checklist of the Preferred Reporting Items Systematic Reviews and Meta-Analyses (PRISMA) standard [21]. In doing so, it is possible to reproduce the processes of searching, selecting, and analyzing the relevant literature. The systematic literature review was conducted in four main phases: identification, selection, eligibility, and inclusion. The flow chart is presented in Figure 1.

Keyword Match. During the identification phase, we performed a structured keyword search using the following sources: ACM, IEEE, and Scopus. Since this paper focuses on cognitive biases related to XAI, the search query was contextualized in three dimensions: *AI systems*, *Explainability*, and *Cognitive biases*. Drawing on the authors' background in XAI, we assigned keywords that describe each dimension. We searched for keywords representing AI systems and Explainability dimensions in the Title, Abstract, and Author Keywords fields, because we wanted to focus on papers whose main topic was XAI. For Cognitive bias keywords, we searched in the Full text of papers. The search result was filtered to include recent papers (2008 or after) since XAI is a young field of study. The search query was as follows, adapted to each database's advanced search specificities (the wildcard * indicates where we retrieved plurals and different spellings):

AI systems ➔ Abstract: (AI, artificial intelligence, machine learning, algorithm*, intelligent system*, neural network*) AND

Explainability ➔ Abstract: (explainab*, explanation*, intelligib*, interpretab*, transparen*, XAI) AND

Cognitive biases ➔ Full Text: (cognitive bias*, decision bias*, explanatory bias*, explanation bias*, human bias*) AND

Date ➔ 2008 and after.

Screening and Eligibility. We considered the following inclusion criteria:

- *Cognitive biases*: The paper describes cognitive biases that are involved in the field of XAI.
- *Mitigation techniques:* The paper describes techniques to mitigate cognitive biases involved in the XAI process.
- *Measurement techniques:* The paper describes ways to measure cognitive biases related to explanations.
- Papers that do not provide primary insights on cognitive bias in XAI are excluded (*e.g.*, a paper that does not provide enough detail on how the heuristics manifest and in what context).

Additionally, only peer-reviewed papers written in English were included. We excluded very few papers to which we did not have access. The identification phase yielded a total of 273 results: 59 papers from ACM, 64 from IEEE, 150 from Scopus, and 12 additional papers selected from the references of relevant papers or based on the authors' knowledge. The authors' names, article title, source title, and publication year of the identified records were exported to an Excel spreadsheet. A total of 261 results were obtained after eliminating 24 duplicates. In the screening stage, each paper's title and abstract was reviewed by an author based on the inclusion and exclusion criteria, and a decision was made as to whether the paper should be rejected or retained for the next phase (eligibility). 176 papers were excluded because they did not discuss cognitive biases involved in the field of XAI. A total of 85 papers were advanced to the next phase. In the eligibility stage, two of the authors read the remaining articles in full. Based on the inclusion and exclusion criteria, a decision was then made as to whether the article should proceed to the final phase. 48 articles were finally excluded at this stage because they did not sufficiently address the proposed research questions (cf. introduction). 37 articles were retained and advanced to the final phase.

Coding book. In the inclusion stage, we started the coding of the papers by having two authors extract relevant information from the papers. Except for the type of article (primary study or survey), this information essentially relates to RQ2 (see introduction). To ensure coding quality, this information was brainstormed by the authors and the research team and was drawn from related surveys of empirical studies of XAI (e.g., [22]). As such, our code book included: Cognitive bias type; Mitigation strategy; Explainability technique and format (local feature explanation, global explanation, etc.); Paper type (primary study or review); Application/domain (high-risk or low risk); AI type (shallow, deep or wizard of oz) and algorithm used (when specified); Human task type (proxy or real and description); Human expertise (lay-user, domain expert or ML expert). The full code description can be found in the appendices.

Corpus presentation. In the corpus of 37 papers we analyzed, 7 papers are reviews of the literature, and 29 papers are primary studies. Figure 2 illustrates the distribution of our corpus across the disciplines, showing the diversity of the subject areas. As we can see, over half of these papers are Human Computer Interaction (HCI) works, published in leading conferences (*e.g.*, CHI and IUI). The remaining papers have also been published in leading conferences and journals directly or indirectly related to the explainability of AI systems, in the fields of AI, computer science and psychology.

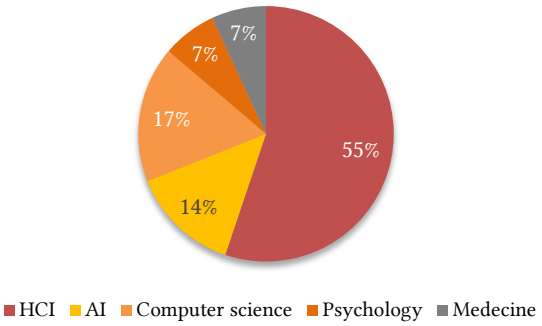

Figure 2: The distribution of our corpus across disciplines

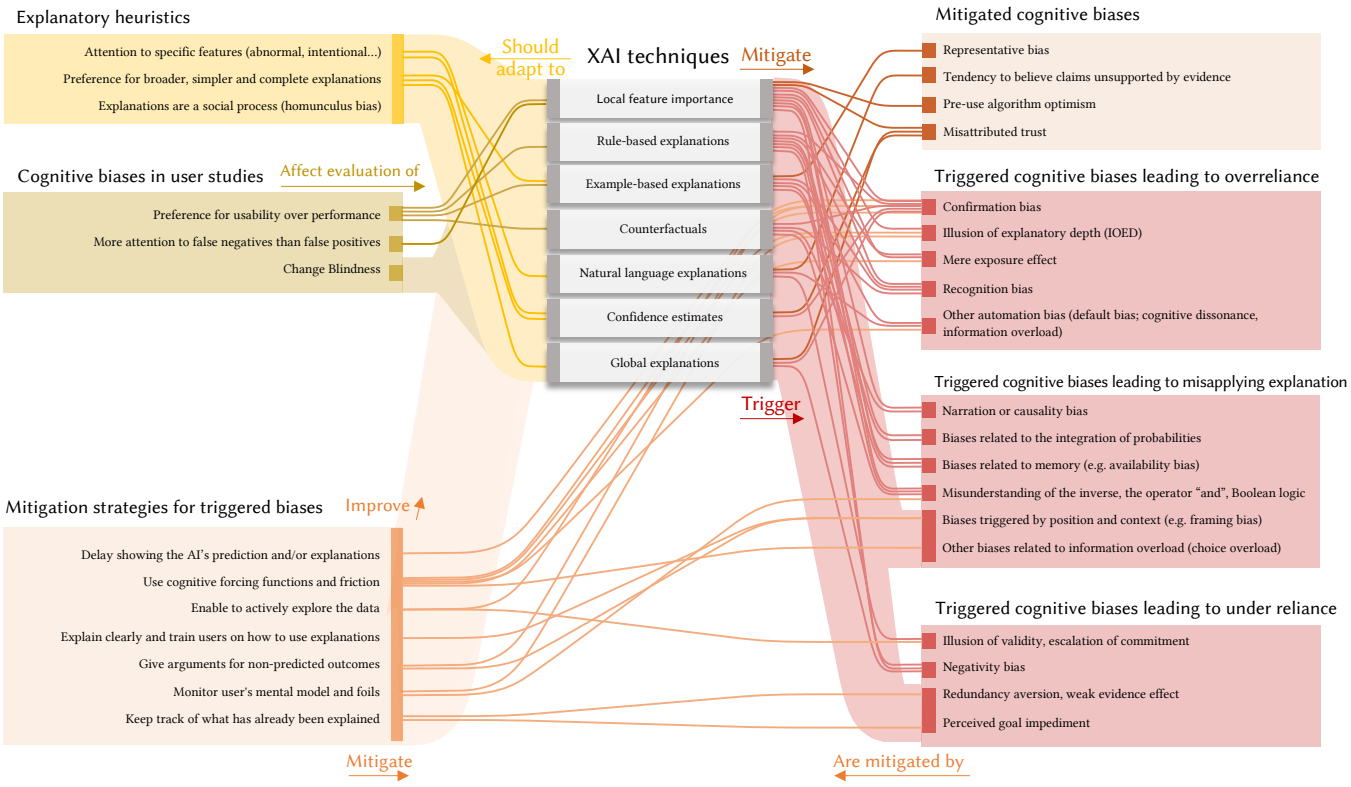

Figure 3: Summarization of the cognitive constraints, biases and mitigation strategies discussed in the papers included in our corpus (n=37). This diagram presents the different categories of explanation techniques that were seen in our corpus (in the middle). Each link represents a connection made in the literature between an explainability technique and a cognitive bias or between a cognitive bias and a mitigation technique. The legends in color underlined by arrows indicate how and in what direction the links should be read (*e.g.* "XAI techniques should adapt to explanatory heuristics"). The pale and wide links indicate that the bias or cognitive constraint applies more generally to all XAI methods. We identified more connections between biases and mitigation strategies but show only the most supported ones for brevity.

Identification of cognitive biases. To identify by name the cognitive effects that were discussed in the papers we reviewed, we either took the wording used in the papers, or if the bias was not named explicitly, we relied on external taxonomies [15], [16], surveys (*e.g.*, [17]), and on our own knowledge of cognitive biases. For a few cases we coined a phrase to be able to refer to the effect under study (*e.g. "pre-use algorithmic optimism"* [23]*).*

3 Results

This section presents the results of the analysis of the articles studied. First, we give an overview of the biases identified (RQ1). We then examine the stated mitigation strategies as well as the research methods used to identify them (RQ2, RQ3 and RQ4). For the sake of brevity, we do not systematically provide the definitions of the biases we examine, but the interested reader can refer to the lexicon provided in the appendices.

3.1 A map of the cognitive biases in XAI

The first contribution of this work is to answer our RQ1 and identify the cognitive biases encountered in our corpus, along with the context in which they were found, namely the explainability technique that was used, the domain, the task, and the user type. We identified a list of 53 cognitive biases (cf. Appendix 1). We then analyzed the way these biases were presented in the articles reviewed, revealing four main ways cognitive biases affect or are affected by XAI-aided human decision systems.

The first type are cognitive biases that affect how XAI methods are designed. They are listed in the yellow boxes in Figure 3 (top-left corner). They include all the explanatory heuristics that people use when explaining or receiving an explanation. These explanatory heuristics are well documented in psychological works on the human explanation process [13], [24]. Unlike the other types of cognitive biases discussed in this article, these explanatory heuristics are not considered to lead to errors. On the contrary, they were simply presented as neither good nor bad but merely constraints to be taken into account before designing explainability techniques.

The second category we identified are cognitive biases which can distort how XAI techniques are evaluated in user studies. They are presented in the brown box in Figure 3 (middle left). Prompted by Doshi-Velez and Kim [25], recent attention has been focused on approaches to evaluating explanations, with some researchers arguing for the need to test explanations with users [26], and

others cautioning against doing so, concerned that cognitive biases could skew evaluations and mislead the XAI field [27]. We take stock of these cognitive biases in section 3.2.2.

The third category are cognitive biases that were successfully mitigated by XAI techniques. They are presented in the dark orange box in Figure 3 (top-right corner). In Section 3.2.3, we review successful examples of using an explainability technique to address a cognitive bias that was observed in an AI-aided decision-making process.

The fourth type of cognitive biases are those caused or exacerbated by explainability, and which lead to erroneous decision-making. They are presented in the red boxes on the right of the diagram in Figure 3. Among these, we find cognitive biases that lead either to overreliance, to under reliance, or to the misapplying of the explanation.

3.2 Contexts in which cognitive biases occur, mitigation and evidencing strategies

Let us go through the different categories of biases identified and answer our research questions RQ2, RQ3 and RQ4, namely in which contexts—explainability technique, user type and task type—those biases were observed, how to mitigate and how to reveal them.

3.2.1 Explanatory heuristics affecting XAI design

In this section we summarize the cognitive biases employed by people to generate, evaluate and communicate explanations. As the term "bias" can bring to mind errors of judgment, we refer to "explanatory heuristics." Unlike the cognitive bias of the other categories in Figure 3, in this class, the explanatory heuristics are inherent to the explanation process and help humans select some events as being relevant 'causes' out of a potentially infinite causal chain of events [28]. In our corpus, "explanatory heuristics" were mainly examined by reviews such as Miller's [13] seminal effort to tie psychological theories on explanations to the field of XAI, but also by primary studies focused on explainability desiderata such as simplicity and completeness.

When generating explanations, attention is drawn to specific causes. Based on [24], [29]–[31] and others, Miller [13] unwinds the cognitive process of explaining. First, people identify candidate causes using abductive reasoning. Second, people use heuristics to select the most relevant among these candidates. The work in our corpus mentioned several features on which people tend to focus when selecting causes. These are abnormality, intentionality and responsibility, necessity, sufficiency and robustness [13], [32], confidence levels [12], [13], [32], demographic features [33], contrast between fact and foil [13], [34], [35] and inherent features [13], [36]. An interesting example of incorporating these attentional biases into XAI techniques is [36], which used the inherence bias—a human tendency to focus on inherent features instead of extrinsic ones to explain a phenomenon—to select explanations for person re-identification systems. Bhatt *et al.* [32] also stress the importance of showing confidence estimates of AI prediction. They argue that people need to assess uncertainty to make decisions, relying on prospect theory [37]. In social interactions, we are used to estimating the confidence level of a person's assertion based on their tone and other social cues. These cues are not applicable in human-IA interactions, hence the need to explicitly state the AI's confidence levels.

When evaluating the quality of an explanation, humans look for specific properties. Existing work on XAI desiderata and psychology of explanations has evidenced that people look for specific qualities in explanations. In our corpus, we also observe such preferences for "broad" [13], [35], [38], "simple" [13], [38]–[40] and "more complete" [41] explanations. However, the preference for simple and complete explanations raises several ambiguities. While it is unchallenged that simpler explanations are more comprehensible and readable [9], [39]—some researchers even show that interpretability is inversely related to explanation length [9]—they can also be received with skepticism by users [9], [41], [42]. In section 3.2.4.2, we discuss how users expect complex concepts to be explained by complex explanations. Similarly, Kulesza *et al.* [41] argue that more comprehensive explanations help to significantly improve participants' mental models, but other work [42], [43] found complete explanations can lead to overreliance [35]. In [35], [38], the authors contend that coherent and broad explanations are preferred, with scope being even more important than simplicity, consistently with Lombrozo's point of view in cognitive science that broader and simpler explanations are better [24]. Based on these findings, it can be challenging to gauge the right level of complexity in explanations. Some suggested general principles such as not providing explanations that are too complex to be readable [9], adjusting to the level of "completeness" to each user and context [35].

Explanations are a social process (homunculus bias). Miller [13] argued that explanations are as much a social process as a cognitive one. The social process allows the explainer to identify the knowledge gap in the recipient's mental model that needs explaining, to refine explanations, use appropriate vocabulary and answer to follow-up questions. Weld and Bansal [34] state that adopting that social process would be highly beneficial to provide more relevant explanations. Moreover, people tend to attribute human traits to machines—known as the homunculus fallacy—and therefore tend to expect that these tools use the same communication framework as humans [13], [34]. To that end, some researchers have argued for more interactive explanations. However, there is some concern in the articles of our corpus that interactive explanations may lead to over reliance. We describe these in section 4.

3.2.2 Cognitive biases arising in user-based evaluations of explainability techniques

Users' stated preferences are not indicative of performance. Buçinca *et al.* [44] warned against using proxy tasks to evaluate explanations through user studies, i.e., tasks consisting of subjectively rating the explanations. They noted that people's subjective preferences for explanations were not indicative of the performance they would exhibit in making decisions with these

explanations. Instead, researchers should use real tasks. This observation was also evidenced in our corpus with local feature importance, rule-based, example-based, and counterfactual explanations [33], [43], [44].

More attention to false negatives than false positives. Focusing on saliency maps for image recognition, Mohseni *et al.* [45] showed that people pay less attention to explanations of false positives than explanations of false negatives. They also showed that people rate differently techniques that differ only in appearance. To address these biases, they designed a human attention baseline to evaluate saliency explanations without having to resort to user studies.

Furthermore, Sokol and Flach [46] called for caution about the phenomenon of **change blindness** in user studies, namely the "inability to notice all of the changes in a presented medium," especially in an image. To address it, any change should be highlighted or made salient. Researchers should also be wary of selection bias when selecting participants for user studies through Amazon Mechanical Turk, usually more computer literate than the 'normal' population [47].

3.2.3 Cognitive biases in AI-aided decisions that are mitigated thanks to explainability

Mitigated bias	Explanation method used to mitigate bias	Task/user
Misattributed trust	Uncertainty estimates [12]	Medical diagnosis / domain expert
Representativeness bias	Prototype cases of decision outcomes [12]	Medical diagnosis / domain expert
Tendency to believe persuasive claims unsupported by evidence	Natural language explanations [48]	Fake news detection / lay users
Pre-use algorithmic optimism	Local feature importance (word highlighting) [23]	Emotional analysis / lay users

Table 1: Cognitive biases mitigated by XAI and the context in which they were revealed.

As Liao *et al.* indicate [49], "users also deem explanations of the AI's decision as potential mitigation of their own decision biases." Here, XAI researchers examine cognitive biases that arise in decision-making, with or without AI systems, and that can be mitigated by explainability. This discussion is usually centered around broad notions of transparency as a potential tool to mitigate aversion bias, see [50] for example. However, we only included in our corpus studies that demonstrated successful mitigation of such biases.

We have found that XAI can mitigate pre-use algorithmic optimism [23], bias towards unsupported persuasive claims [48], representativeness bias through the means of example-based explanations [12] and misattributed trust through carefully designed uncertainty estimates [12]. For the sake of brevity, we discuss only one of these examples below. Interested readers may refer to the citations in Table 1 for more information on these examples.

Pre-use algorithmic optimism. Springer and Whittaker [23] evidence how users had positive expectations of the transparent system before using it. To be able to refer to it later, we call this phenomenon "pre-use algorithmic optimism." Springer and Whittaker conclude that showing explanations, in this case local feature importance, was important to prevent users from overestimating the capabilities of the system. They suggest presenting explanations gradually or only when requested, to prevent users from losing trust when their expectations about the system are contradicted.

3.2.4 Cognitive biases in AI-aided decisions that are exacerbated by explainability

Recently, there have been concerns that AI explanations can bias users and impair their decision-making process [11], [51], [52]. At the root of this issue, Buçinca *et al.* [44] argue, is the choice between trusting an AI recommendation or engaging in an effortful and time consuming cognitive analysis of its explanations. People thus "develop heuristics about whether and when to follow the AI suggestions" [44], and AI explanations can reinforce such heuristics. However, as pointed out by [51], current explanation techniques do not provide causal guarantees, which means they can be wrong, adding another potential source of error in addition to the AI system. In such a situation, it becomes very difficult for humans to detect errors and "intervene meaningfully" [53], also considering that human performance in deception detection is quite poor as shown in [54] in a task of detecting deceptive hotel reviews. Table 2 summarizes the different cognitive biases that were presented as caused or enhanced by explanations of AI recommendations in the articles we reviewed. We classified them in three categories based on the type of error they result in: misapplying explanations, overreliance and under reliance. For brevity, we do not discuss every cognitive bias, associated context, evidencing and mitigation strategies but the reader can refer to the lexicon in the Appendix 1 and Table 3 for additional details.

3.2.4.1 Explanation-triggered cognitive biases that lead to misapplying the explanation

'Narrative' or 'causal' bias. A quote from Tversy and Kahneman about causal bias [16] applies wonderfully to the context of XAI: *"in the context of explanation and revision, the strength of causal reasoning and the weakness of diagnostic reasoning are manifest in the great ease with which people construct causal accounts for outcomes which they could not predict".* Several articles in our corpus emphasized that experts were particularly affected by this heuristic, including researchers who attribute causal learning to saliency maps [55], data scientists who make false narratives about how SHAP and GAM explanations work [52] or domain experts in the domain of child welfare screening using counterfactuals [40]. The authors mainly called for incorporating knowledge-based narratives in explanations and [55] encouraged researchers to use direct experimental evidence to back up their claims.

Related to the integration of probabilities. In their review of biases related to rule-based explanations, Kliegr *et al.* [17]

described several cognitive biases related to people's difficulty to integrate probabilities such as base rate neglect or conjunction fallacy. Fürnkranz *et al.* [9] further evidenced that people (lay users in this case) tend to ignore the statistical significance of a statement, a phenomenon called insensitivity to sample size. Miller [13] stressed that probabilities don't matter to people – a claim somewhat disputed by [32] if uncertainty estimates are probabilities (cf. section 3.2.1.1) - and that explanations should focus on causal relationships. Others showed that some user's individual characteristics impact the way explanations are received. Coba [56] used a Choice-Based Methodology [57] and eye-tracking measurements to reveal that people's various decision making styles impact how they perceive hotel ratings—shown as "collaborative explanations". People of the "maximizer" type were more prone to insensitivity to sample variance and choice overload.

Related to memory: Wang *et al.* [12] discussed representativeness and availability bias in the context of medical diagnosis, and proposed showing prior probability and prototypes of outcomes to mitigate these.

Biases leading to misusing the explanations can also be due to **misunderstanding some elements of the language** [17] that is commonly used in explanations such as the logical operator "and" in rules [9], Boolean logic in counterfactuals [40], or confidence scores when it is ambiguous what they refer to [32].

Related to position and context [17]**.** For example, Nourani *et al.* [11] discuss the primacy effect, a tendency to form an opinion based solely on the first piece of information received. They suggest controlling the type of predictions users observe when first interacting with the system.

3.2.4.2 Explanation-triggered cognitive biases that lead to overreliance

According to the **mere exposure effect** [17], the sheer presence of an explanation increases confidence in the machine's prediction. This effect was evidenced by [8], [9], [54], with lay users, rule-based and local feature importance explanations, by demonstrating that random or placebic explanations increase trust. Several papers examined **user's bias for completeness** [9], [41]–[43], [54]. For example, Fürnkranz *et al.* [9] showed that users found longer explanations more plausible than shorter ones. This is consistent with [42] which showed that giving a fuller explanation in the context of a medical diagnosis led to over-reliance issues, with [54], which demonstrated that additional details including irrelevant ones improved user's trust in AI predictions, and with [43] which contends that the additional details contained in visual explanations compared to textual ones increase users' misattributed trust. Szymanski et al. [43] showed that lay users were more exposed to confirmation and completeness bias than machine learning experts when faced with visual explanations of a reading time prediction algorithm. These articles provide several avenues for addressing this problem, including by combining the use of textual and visual explanations [43] or by providing arguments against the machine's suggestion [42]. Some mentioned the possibility that more complete explanations are more likely to contain elements that the user recognizes, thus contributing to the persuasive effect through the **recognition bias** [17]. Another bias studied in the corpus is the phenomenon called **"illusion of explanatory depth,"** coined by Koehler [58] and evidenced in the XAI literature by Chromik *et al.* [10] using local feature importance (SHAP [1]) explanations. They prompted users to self-explain so that they would realize that they knew less about the concept being explained than they had originally imagined. We can also perceive this effect in [52], [59] which mention "superficial" and "rush" understanding.

3.2.4.3 Explanation-triggered cognitive biases that lead to under reliance

Our corpus also contains articles discussing under reliance issues. These were manifested through various effects, such as **"the escalation of commitment"** [60], the **"illusion of validity"** [61] or **"perceived goal impediment"** [59], which concerned mainly domain experts, and **"negativity bias"** [11], [17], [38], [40], [62]. Several works have highlighted the role of user expertise in under reliance problems. Domain experts have developed cognitive routes that enable them to make quick and accurate decisions in environments that are "regular" enough to be predictable [63]. Their intuition is therefore more sophisticated than a lay user's "System 1"[15]. Simkute *et al.* [61] highlight Klein's [64] results, indicating that experts make decisions intuitively, with little uncertainty, and rarely consider more than one option. While useful heuristics, these reasonings also make them more prone to belief perseverance [58] or algorithmic aversion [65], especially when faced with contradictions from the machine's predictions [61]. In addition, user studies involving domain experts reproduce decision-making conditions that are representative of real-world situations—sometimes high-stakes and time-limited, therefore more stressful —which may explain the reluctance of experts to engage in explanations. Negativity bias, a tendency to pay more attention to negative features of the AI or AI explanations, was found to affect everyone including non-expert users. Nourani *et al.* [11] suggest controlling what types of predictions users see when first interacting with the system. For example, showing the weaknesses of the system early on will have a major influence on trust, as will showing negative outcomes early on, such as a malignant diagnosis.

4 Discussion: A research agenda to address cognitive biases in XAI

Based on the findings of the articles we reviewed, the methods they used to expose cognitive biases, and the mitigation strategies they outlined, we present below a discussion of research directions we believe should be pursued in future work to address cognitive biases in XAI.

Clarify the "normal" vs. "problematic" cognitive biases. Which cognitive biases need to be mitigated? In this paper, we identified some cognitive biases as being neutral heuristics, *i.e.*

Cognitive biases	Examples of evidencing strategies (in user studies)	Examples of mitigating strategies	Ref. in corpus
Caused, triggered, or enhanced by XAI leading to misapplying the explanation			
Related to causality: Narrative bias, Over-generalization, Causation vs. correlation, attention to demographic features	Ask participants to describe explanations, analyze free text answers and verbalizations [52].	Incorporate human expertise into explanations [67].	[40], [52], [55], [67]
Related to the integration of probabilities: Averaging bias, Base-rate neglect, Conjunction fallacy, Disjunction fallacy, Insensitivity to sample size, Unit bias	Measure the correlation between the user's confidence and supporting evidence [9], [56].	Reminder of probability theory. Use frequencies instead of percentages. Show support as an absolute number [17].	[9], [17], [56]
Related to memory: Representativeness, Availability bias	Analyze reasoning process through free text questions and think-aloud protocols [12], [40].	Show prior probabilities of outcome and examples of decision outcome [12].	[9], [12], [17], [40], [52]
Triggered by misunderstanding of language: Misunderstanding of the inverse, of 'and', Boolean logic, confidence scores	Analyze free text responses [40], Clarify the meaning of language elements to only one group of participants [9].	Clearly communicate what the presented information means [42]. State only true statements for the presentation of Boolean elements, including by negating false ones [40]	[9], [17], [40], [42]
Triggered by position and context: Framing bias, Primacy effect, Anchoring bias	Measure the perceived reasonableness of explanations and the performance of users at a task under different explanation framing conditions [11], [68].	Describe the uncertainty of both positive and negative outcomes [32]. Control the kind of predictions users observe in the training phase [11].	[11], [12], [17], [32], [34], [59], [68]
Related to information overload Choice overload	Measure the user's cognitive load using the NASA Task Load Index (NASA-TLX) [23], [52], Eye-tracking measurements (for choice overload) [56]	Do not use too many explainability types [40]. Use user-centric approaches [59].	[40], [59], [61], [69]
Caused, triggered, or enhanced by XAI leading to overreliance			
Completeness bias, Cognitive dissonance, Confirmation bias, Default bias, Illusion of explanatory depth, Mere exposure effect, Other automation bias, Recognition bias	Observe user's degree of agreement with the AI and user's comments with vs. without explanations [48]. Study the correlation between explanation length and perceived plausibility [9].	Give arguments for non-predicted outcomes [12], [34], [42]. Delay showing the AI's prediction and/or explanations [12], [23], [44], [54]. Use cognitive forcing functions and friction [44], [61], [69]. Include uncertainty estimates [12], [32], [42].	[8]–[10], [12], [17], [33], [35], [41]–[43], [48], [52], [54], [56], [59], [60], [69]
Caused, triggered, or enhanced by XAI leading to under reliance			
Escalation of commitment, Illusion of validity, Negativity bias, Familiarity bias, Perceived goal impediment, Redundancy aversion, Weak evidence effect	Observe the relation between subjective confidence, subjective comprehension, and positive and negative AI outcomes [11]. Ask participants to think aloud while they make decisions [12].	Enable to actively explore the data [12], [61]. Use gamification and personalization [61]. Keep track of what has already been explained [13], [69]. Control the predictions users observe in the training phase [11].	[9], [11], [17], [38], [40], [44], [59]–[62], [69]

Table 2: Cognitive biases triggered or exacerbated by XAI and the context in which they were evidenced.

"normal" ones, inherent to the process of explanations. Instead of mitigating those biases, some [13], [34] argue that they should be taken into account in the design of explanations, for example by providing explanations as social processes or by adopting contrastive explanations. However, there is a blurred line between biases XAI needs to adapt to and those that need to be mitigated. It goes back to the important question posed by Weld and Bansal [34]: "Should an explanation system exploit human limitations or seek to protect us from them?". Lakkaraju *et al.* [66] argue that by exploiting certain human cognitive biases, such as preferences for relevant or familiar features, trust could be manipulated. Conversely, Miller argues explanations should be contrastive, simple and when applicable delivered in the form of a dialogue, *i.e.* interactive [13]. Clarifying which biases are normal and which are undesirable appears to be important for moving the XAI field forward. To that end, more empirical work on the benefits and drawbacks of incorporating cognitive constraints into explanation is needed. Specifically, future work could investigate on some currently puzzling results [23]. For instance, there has been a surge of interest in interactive explanations recently, responding to the call to design explanations that fit the social process of explanation [34]. However, concerns were expressed in [33] as interactive explanations were found to reinforce user's over reliance on AI suggestions. A possibility is that interactive explanations were more complex to interpret in [33]'s study, leading to information overload. More work is still needed on the correct calibration of such interactive explanations. We also found contradictory results between Zytek *et al.* [40] which found that

example-based explanations for child welfare screening led to representativeness bias and Wang *et al.* [12] which presented prototypes of decision outcomes as a mitigation for the same bias. In addition, Lai and Tan [54] warned to be cautious about the "backfire effect" according to which "corrections of misperceptions may enhance people's false beliefs" [70]. Kliegr *et al.* [17] also mentioned the possibility that different cognitive biases could have opposing effects, such as information bias (leading to overreliance) and ambiguity aversion (leading to under reliance), thus emphasizing the need to consider biases in their context and to put them in relation to the user's knowledge.

More normative work on assessments of XAI systems. As the foregoing discussion highlights, there needs to be not only more empirical research on bias, but also more theoretical and normative work to distinguish between processes that are truly biased, *i.e.,* distorted and in need of modification, and those that are "normal." Such a distinction seems difficult to make without normative evaluations referring to the correctness of decisions and the inherent quality of the decision process for the users, including his or her level of participation. Work to identify the normative seriousness of various biases could help researchers and XAI designers decide whether, how and in which priority different biases need to be addressed, as well as make relevant tradeoffs more explicit.

Complete frameworks of stakeholders in XAI. To meet the cognitive needs of the user, the idea of tailoring explanations to the task at hand, to the user's goals, knowledge [33], [35], [43], [56] and her specific needs (such as exploring the raw data for experts [12], [61]) currently fails to take biases into account. Future work could address how to complete the detailed taxonomies of users' expertise and their role in XAI [71]–[73] by considering the cognitive biases they may be prone to.

Improve the perception of the user's reactions to XAI. Several authors have advocated that we need a better perception of social and emotional behavior of users to be able to correct errors in their reasoning and their mental models of the system [10], [35], [74]. As a first step towards this, we highlighted some methods to evidence biases in Table 2. Notably, what seems to be a good practice for controlling for the mere exposure effect is using placebic explanations or randomly generated explanations as a baseline [8], [11]. Then, cognitive load can be measured through the means of the TLX workload assessment method [23], [52], eye-tracking measurements [56] or through the number of cognitive chunks and a subjective measure encompassing the reading time, the self-reported load and memory performance (how well the user remembers the explanation) [39]. In addition, we frequently encountered the use of qualitative analyses in our review, such as think-aloud protocols [12], [23], [43], [59], useful as pre-studies but not generalizable (they involved from 12 to 20 participants in our corpus), or the analysis of free text comments, which can be implemented more easily on a larger scale [40], [43], [69]. Further, the ability of XAI systems to capture users' mental states could be complemented by a memory of these states and a memory of what has already been explained [13], [59].

Evaluate XAI techniques with human attention, without human biases. We have outlined in section 3.2.4 a few cognitive biases that can skew evaluations of explainability approaches. However, solely relying on quantitative metrics for the evaluation of explanations could lead to overlooking essential considerations regarding the human user. To address this, Mohseni *et al.* [45] presented a promising evaluation methodology, without humans, but taking into account human attention. Leveraging human annotators, they developed human attention masks which can be used to evaluate model saliency explanations for image and text domains. Future research could continue in that line of work. Meanwhile, quantitative based experiments such as the sanity checks for saliency maps performed in [75] are equally important to criticize existing techniques [76].

Work *around* explanations. Various work in our corpus mentioned the need to pay more attention to other interaction design choices [44], [77] beyond the choice of an explanation method. These include contextual information, training, timing, framing, and other specific strategies to mitigate cognitive biases. For example, Simkute *et al.* [61] suggested the use of gamification strategies in low-stakes environments to address the lack of motivation of some users, and the use of feedback and controls in high-stakes environments. Others stressed the need to clarify specific elements in the explanations. Bussone *et al.* [42] proposed presenting how the explanations were derived, which Dazelay [78] calls 'meta-explanations'. [12], [44], [54] suggested to delay showing the AI's prediction and/or explanations to decrease overreliance issues. Nourani *et al.* [11] recommended to control the type of predictions that users observe when learning to use the system, during the initial instructions and training phase. Finally, [44], [59] proposed cognitive forcing functions and friction-based strategies to address users' lack of curiosity. Cognitive forcing functions consisted in making users wait for the explanations, updating them or asking for them. The friction function designed by Naiseh consisted in asking the user to confirm they did not want to review the explanation. All these strategies proved to be useful in decreasing user's unjustified trust, though it decreased their satisfaction in the system.

Give arguments against the prediction. The idea of explaining not only the AI's prediction but also alternative possibilities appeared in several papers [12], [34], [42] as a way to counter automation bias. Wang *et al.* recommended to support "premortem of decision outcomes", a reasoning consisting in trying to disprove a hypothesis. Bussone *et al.* [42] highlighted comments from participants saying they wanted to see both positive and negative evidence for the suggested medical diagnosis. Finally, Bansal *et al.* [79] envisioned an AI that would play *"a devil's advocate role, explaining its doubts, even when it agrees with the human."* They proposed a prototype of such an explanation and found that while it was effective in informing the human that the AI might be wrong, it was not sufficient to reduce

significantly errors related to overreliance. One of the main challenges is getting users to come up with their own solution when they are informed that the AI may be wrong. Additional work is still needed to find the right kind of interaction that could help users detect that the AI is wrong [79], but the direction seems promising, notably for two reasons. First, it reminds us of the adversarial structure of a judicial system where two parties (a defense attorney and a prosecutor) present opposing arguments. Implementing such "adversarial explanations" could increase societal trust in the AI-aided decision process. Second, a necessary condition for free will is the availability of alternative possibilities, or the ability to "choose otherwise" [80]. Therefore, showing alternative explanations to the decision-maker helps with sustaining her autonomy and accountability.

Limitations

Since our goal was to provide insight into how the XAI field has considered cognitive biases to date, we used a systematic search methodology. This allowed us to cover a broad sample of articles on XAI. However, it is possible that some articles did not use our general search terms on cognitive biases and focused on specific types of cognitive biases in XAI. Our paper augmentation is limited by potential biases in the authors' view of the XAI field. To continue this line of research on cognitive biases, future review work could focus on specific biases, such as "automation bias". Evidently, our list of 53 cognitive biases cannot be considered as the finite list of biases affecting XAI systems, there are numerous others in the cognitive science literature which may be worth studying in the context of XAI. Moreover, it was quite difficult to assess the generalizability of the results presented in our corpus. To address this limitation, we tried to preserve the context in which these results were obtained—explainability technique, user type, and task type. However, it is possible that these results depend on more granular details. Finally, we leave it for future work to produce more interactive versions of a heuristic map such as the one we present, in a similar fashion as Suresh *et al.* [73]. This could facilitate the tracking of cognitive biases that have been highlighted in the XAI literature and the contexts in which they have been highlighted.

Conclusion

In this paper, we presented a systematic review of 37 papers—drawn from a corpus of 285 papers—to investigate what kind of cognitive biases were identified in the presence of XAI systems. In addition, we carried out a qualitative analysis of these papers, providing a map of the different cognitive biases and revealing in which context they occur, for example with which XAI techniques, which type of users and AI-assisted task. Furthermore, our mapping reveals the different ways these biases affect XAI-aided decisions: 1) cognitive biases affect how XAI methods are designed, 2) they can distort how XAI techniques are evaluated in user studies, 3) some cognitive biases can be successfully mitigated by XAI techniques, and, on the contrary, 4) some cognitive biases can be exacerbated by XAI techniques. Finally, we provide several directions for future work that pave the way for meeting users' cognitive needs, which is an important development towards a human-centered XAI.

REFERENCES

[1] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, Dec. 2017, pp. 4768–4777.

[2] M. T. Ribeiro, S. Singh, and C. Guestrin, "'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*, New York, NY, USA, Aug. 2016, pp. 1135–1144.

[3] A. Ghandeharioun, B. Kim, C.-L. Li, B. Jou, B. Eoff, and R. W. Picard, "DISSECT: Disentangled Simultaneous Explanations via Concept Traversals," *ArXiv210515164 Cs*, Feb. 2022.

[4] B. Kim, C. Rudin, and J. Shah, "The Bayesian Case Model: A Generative Approach for Case-Based Reasoning and Prototype Classification," p. 9, 2015.

[5] H. Lakkaraju, E. Kamar, R. Caruana, and J. Leskovec, "Interpretable & Explorable Approximations of Black Box Models," *ArXiv170701154 Cs*, Jul. 2017.

[6] A. Lucic, H. Haned, and M. de Rijke, "Why does my model fail? contrastive local explanations for retail forecasting," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, New York, NY, USA, Jan. 2020, pp. 90–98.

[7] M. T. Ribeiro, S. Singh, and C. Guestrin, "Anchors: High-Precision Model-Agnostic Explanations," *Proc. AAAI Conf. Artif. Intell.*, vol. 32, no. 1, Art. no. 1, Apr. 2018.

[8] M. Eiband, D. Buschek, A. Kremer, and H. Hussmann, "The Impact of Placebic Explanations on Trust in Intelligent Systems," in *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA, May 2019, pp. 1–6.

[9] J. Fürnkranz, T. Kliegr, and H. Paulheim, "On cognitive preferences and the plausibility of rule-based models," *Mach. Learn.*, vol. 109, no. 4, pp. 853–898, Apr. 2020.

[10] M. Chromik, M. Eiband, F. Buchner, A. Krüger, and A. Butz, "I Think I Get Your Point, AI! The Illusion of Explanatory Depth in Explainable AI," in *26th International Conference on Intelligent User Interfaces*, New York, NY, USA, Apr. 2021, pp. 307–317.

[11] M. Nourani *et al.*, "Anchoring Bias Affects Mental Model Formation and User Reliance in Explainable AI Systems," in *26th International Conference on Intelligent User Interfaces*, New York, NY, USA, Apr. 2021, pp. 340–350.

[12] D. Wang, Q. Yang, A. Abdul, and B. Y. Lim, "Designing Theory-Driven User-Centric Explainable AI," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA, May 2019, pp. 1–15.

[13] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artif. Intell.*, vol. 267, pp. 1–38, Feb. 2019.

[14] Z. C. Lipton, "The mythos of model interpretability," *Commun. ACM*, vol. 61, no. 10, pp. 36–43, Sep. 2018.

[15] D. Kahneman, *Thinking, fast and slow*. New York, NY, US: Farrar, Straus and Giroux, 2011, p. 499.

[16] D. Kahneman, S. P. Slovic, P. Slovic, A. Tversky, and C. U. Press, *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge University Press, 1982.

[17] T. Kliegr, Š. Bahník, and J. Fürnkranz, "A review of possible effects of cognitive biases on interpretation of rule-based machine learning models," *Artif. Intell.*, vol. 295, p. 103458, Jun. 2021.

[18] C. Rastogi, Y. Zhang, D. Wei, K. R. Varshney, A. Dhurandhar, and R. Tomsett, "Deciding Fast and Slow: The Role of Cognitive Biases in AI-assisted Decision-making," *ArXiv201007938 Cs*, Oct. 2020.

[19] B. Green and Y. Chen, "The Principles and Limits of Algorithm-in-the-Loop Decision Making," *Proc. ACM Hum.-Comput. Interact.*, vol. 3, no. CSCW, p. 50:1-50:24, Nov. 2019.

[20] B. Mittelstadt, C. Russell, and S. Wachter, "Explaining Explanations in AI," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, New York, NY, USA, Jan. 2019, pp. 279–288.

[21] D. Moher, A. Liberati, J. Tetzlaff, D. G. Altman, and The PRISMA Group, "Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement," *PLoS Med.*, vol. 6, no. 7, p. e1000097, Jul. 2009.

[22] V. Lai, C. Chen, Q. V. Liao, A. Smith-Renner, and C. Tan, "Towards a Science of Human-AI Decision Making: A Survey of Empirical Studies," *ArXiv211211471 Cs*, Dec. 2021.

[23] A. Springer and S. Whittaker, "Progressive disclosure: empirically motivated approaches to designing effective transparency," in *Proceedings of the 24th International Conference on Intelligent User Interfaces*, New York, NY, USA, Mar. 2019, pp. 107–120.

[24] T. Lombrozo, "Simplicity and probability in causal explanation," *Cognit. Psychol.*, vol. 55, no. 3, pp. 232–257, Nov. 2007.

[25] F. Doshi-Velez and B. Kim, "Towards A Rigorous Science of Interpretable Machine Learning," *ArXiv170208608 Cs Stat*, Mar. 2017.

[26] F. Poursabzi-Sangdeh, D. G. Goldstein, J. M. Hofman, J. W. Vaughan, and H. Wallach, "Manipulating and Measuring Model Interpretability," *ArXiv180207810 Cs*, Nov. 2019.

[27] B. Herman, "The Promise and Peril of Human Evaluation for Model Interpretability," *ArXiv171107414 Cs Stat*, Oct. 2019.

[28] D. J. Hilton, "Logic and causal attribution," in *Contemporary science and natural explanation: Commonsense conceptions of causality*, New York, NY, US: New York University Press, 1988, pp. 33–65.

[29] T. Lombrozo, "The structure and function of explanations," *Trends Cogn. Sci.*, vol. 10, no. 10, pp. 464–470, Oct. 2006.

[30] B. F. Malle, *How the Mind Explains Behavior: Folk Explanations, Meaning, and Social Interaction*. Cambridge, MA, USA: A Bradford Book, 2004.

[31] M. M. A. de Graaf and B. F. Malle, "How People Explain Action (and Autonomous Intelligent Systems Should Too)," presented at the 2017 AAAI Fall Symposium Series, Oct. 2017.

[32] U. Bhatt *et al.*, "Uncertainty as a Form of Transparency: Measuring, Communicating, and Using Uncertainty," in *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, New York, NY, USA, Jul. 2021, pp. 401–413.

[33] H. Liu, V. Lai, and C. Tan, "Understanding the Effect of Out-of-distribution Examples and Interactive Explanations on Human-AI Decision Making," *Proc. ACM Hum.-Comput. Interact.*, vol. 5, no. CSCW2, p. 408:1-408:45, Oct. 2021.

[34] D. S. Weld and G. Bansal, "The challenge of crafting intelligible intelligence," *Commun. ACM*, vol. 62, no. 6, pp. 70–79, May 2019.

[35] C. Woodcock, B. Mittelstadt, D. Busbridge, and G. Blank, "The Impact of Explanations on Layperson Trust in Artificial Intelligence-Driven Symptom Checker Apps: Experimental Study," *J. Med. Internet Res.*, vol. 23, no. 11, p. e29386, Nov. 2021.

[36] E. Bekele, W. E. Lawson, Z. Horne, and S. Khemlani, "Implementing a Robust Explanatory Bias in a Person Re-identification Network," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Jun. 2018, pp. 2246–22467.

[37] D. Kahneman and A. Tversky, "Prospect Theory: An Analysis of Decision under Risk," *Econometrica*, vol. 47, no. 2, pp. 263–291, 1979.

[38] A. Shimojo, K. Miwa, and H. Terai, "How Does Explanatory Virtue Determine Probability Estimation?—Empirical Discussion on Effect of Instruction," *Front. Psychol.*, vol. 11, 2020.

[39] A. Abdul, C. von der Weth, M. Kankanhalli, and B. Y. Lim, "COGAM: Measuring and Moderating Cognitive Load in Machine Learning Model Explanations," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA, Apr. 2020, pp. 1–14.

[40] A. Zytek, D. Liu, R. Vaithianathan, and K. Veeramachaneni, "Sibyl: Understanding and Addressing the Usability Challenges of Machine Learning In High-Stakes Decision Making," *ArXiv210302071 Cs*, Sep. 2021.

[41] T. Kulesza, S. Stumpf, M. Burnett, S. Yang, I. Kwan, and W. Wong, "Too much, too little, or just right? Ways explanations impact end users' mental models," in *2013 IEEE Symposium on Visual Languages and Human Centric Computing*, Sep. 2013, pp. 3–10.

[42] A. Bussone, S. Stumpf, and D. O'Sullivan, "The Role of Explanations on Trust and Reliance in Clinical Decision Support Systems," in *2015 International Conference on Healthcare Informatics*, Oct. 2015, pp. 160–169.

[43] M. Szymanski, M. Millecamp, and K. Verbert, "Visual, textual or hybrid: the effect of user expertise on different explanations," in *26th International Conference on Intelligent User Interfaces*, College Station TX USA, Apr. 2021, pp. 109–119.

[44] Z. Buçinca, M. B. Malaya, and K. Z. Gajos, "To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making," *Proc. ACM Hum.-Comput. Interact.*, vol. 5, no. CSCW1, p. 188:1-188:21, Apr. 2021.

[45] S. Mohseni, J. E. Block, and E. Ragan, "Quantitative Evaluation of Machine Learning Explanations: A Human-Grounded Benchmark," in *26th International Conference on Intelligent User Interfaces*, College Station TX USA, Apr. 2021, pp. 22–31.

[46] K. Sokol and P. Flach, "Explainability fact sheets: a framework for systematic assessment of explainable approaches," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, Barcelona Spain, Jan. 2020, pp. 56–67.

[47] N. M. Barbosa and M. Chen, "Rehumanized Crowdsourcing: A Labeling Framework Addressing Bias and Ethics in Machine Learning," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA, May 2019, pp. 1–12.

[48] V. Danry, P. Pataranutaporn, Y. Mao, and P. Maes, "Wearable Reasoner: Towards Enhanced Human Rationality Through A Wearable Device With An Explainable AI Assistant," in *Proceedings of the Augmented Humans International Conference*, New York, NY, USA, Mar. 2020, pp. 1–12.

[49] Q. V. Liao, D. Gruen, and S. Miller, "Questioning the AI: Informing Design Practices for Explainable AI User Experiences," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA, Apr. 2020.

[50] H. Park, D. Ahn, K. Hosanagar, and J. Lee, "Human-AI Interaction in Human Resource Management: Understanding Why Employees Resist Algorithmic Evaluation at Workplaces and How to Mitigate Burdens," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA, May 2021, pp. 1–15.

[51] M. Ghassemi, L. Oakden-Rayner, and A. L. Beam, "The false hope of current approaches to explainable artificial intelligence in health care," *Lancet Digit. Health*, vol. 3, no. 11, pp. e745–e750, Nov. 2021.

[52] H. Kaur, H. Nori, S. Jenkins, R. Caruana, H. Wallach, and J. Wortman Vaughan, "Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning," *Proc. 2020 CHI Conf. Hum. Factors Comput. Syst.*, pp. 1–14, Apr. 2020.

[53] European Commission, "Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act)," Apr. 2021.

[54] V. Lai and C. Tan, "On Human Predictions with Explanations and Predictions of Machine Learning Models: A Case Study on Deception Detection," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, New York, NY, USA, Jan. 2019, pp. 29–38.

[55] A. Atrey, K. Clary, and D. Jensen, "Exploratory Not Explanatory: Counterfactual Analysis of Saliency Maps for Deep Reinforcement Learning," *ArXiv191205743 Cs*, Feb. 2020.

[56] L. Coba, L. Rook, M. Zanker, and P. Symeonidis, "Decision making strategies differ in the presence of collaborative explanations: two conjoint studies," in *Proceedings of the 24th International Conference on Intelligent User Interfaces*, New York, NY, USA, Mar. 2019, pp. 291–302.

[57] J. J. Louviere, T. N. Flynn, and R. T. Carson, "Discrete Choice Experiments Are Not Conjoint Analysis," *J. Choice Model.*, vol. 3, no. 3, pp. 57–72, Jan. 2010.

[58] D. J. Koehler, *Explanation, Imagination, and Confidence in Judgment*. 1991.

[59] M. Naiseh, D. Cemiloglu, D. Al Thani, N. Jiang, and R. Ali, "Explainable Recommendations and Calibrated Trust: Two Systematic User Errors," *Computer*, vol. 54, no. 10, pp. 28–37, Oct. 2021.

[60] S. Bayer, H. Gimpel, and M. Markgraf, "The role of domain expertise in trusting and following explainable AI decision support systems," *J. Decis. Syst.*, vol. 0, no. 0, pp. 1–29, Aug. 2021.

[61] A. Simkute, E. Luger, M. Evans, and R. Jones, "Experts in the Shadow of Algorithmic Systems: Exploring Intelligibility in a Decision-Making Context," in *Companion Publication of the 2020 ACM Designing Interactive Systems Conference*, New York, NY, USA, Jul. 2020.

[62] D. Branley-Bell, R. Whitworth, and L. Coventry, "User Trust and Understanding of Explainable AI: Exploring Algorithm Visualisations and User Biases," in *Human-Computer Interaction. Human Values and Quality of Life: Thematic Area, HCI 2020, Held as Part of the 22nd International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings, Part III*, Berlin, Heidelberg, Jul. 2020, pp. 382–399.

[63] D. Kahneman and G. Klein, "Conditions for intuitive expertise: A failure to disagree," *Am. Psychol.*, vol. 64, no. 6, pp. 515–526, 2009.

[64] G. A. Klein, *Sources of Power: How People Make Decisions*. Nature, 1988.

[65] B. J. Dietvorst, J. P. Simmons, and C. Massey, "Algorithm aversion: People erroneously avoid algorithms after seeing them err.," *J. Exp. Psychol. Gen.*, vol. 144, no. 1, pp. 114–126, 2015.

[66] H. Lakkaraju and O. Bastani, "'How do I fool you?': Manipulating User Trust via Misleading Black Box Explanations," *Proc. AAAIACM Conf. AI Ethics Soc.*, pp. 79–85, Feb. 2020.

[67] N. Andrienko, G. Andrienko, L. Adilova, S. Wrobel, and T.-M. Rhyne, "Visual Analytics for Human-Centered Machine Learning," *IEEE Comput. Graph. Appl.*, vol. 42, no. 1, pp. 123–133, Feb. 2022,

[68] T. Kim and H. Song, "The Effect of Message Framing and Timing on the Acceptance of Artificial Intelligence's Suggestion," in *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA, Apr. 2020, pp. 1–8.

[69] M. Naiseh, R. S. Al-Mansoori, D. Al-Thani, N. Jiang, and R. Ali, "Nudging through Friction: An Approach for Calibrating Trust in Explainable AI," in *2021 8th International Conference on Behavioral and Social Computing (BESC)*, Oct. 2021, pp. 1–5.

[70] B. Nyhan and J. Reifler, "When corrections fail: The persistence of political misperceptions," *Polit. Behav.*, vol. 32, no. 2, pp. 303–330, 2010,

[71] R. Tomsett, D. Braines, D. Harborne, A. Preece, and S. Chakraborty, "Interpretable to Whom? A Role-based Model for Analyzing Interpretable Machine Learning Systems," *ArXiv180607552 Cs*, Jun. 2018.

[72] S. Mohseni, N. Zarei, and E. D. Ragan, "A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems," *ArXiv181111839 Cs*, Aug. 2020.

[73] H. Suresh, S. R. Gomez, K. K. Nam, and A. Satyanarayan, "Beyond Expertise and Roles: A Framework to Characterize the Stakeholders of Interpretable Machine Learning and their Needs," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA, May 2021, pp. 1–16.

[74] Z. Akata *et al.*, "A Research Agenda for Hybrid Intelligence: Augmenting Human Intellect With Collaborative, Adaptive, Responsible, and Explainable Artificial Intelligence," *Computer*, vol. 53, no. 8, pp. 18–28, Aug. 2020.

[75] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity Checks for Saliency Maps," *ArXiv181003292 Cs Stat*, Nov. 2020.

[76] B. Kim, R. Khanna, and O. Koyejo, "Examples are not Enough, Learn to Criticize! Criticism for Interpretability," in *Advances in Neural Information Processing Systems 29 (NeurIPS 2016)*, 2016.

[77] Z. T. Zhang, Y. Liu, and H. Hussmann, "Forward Reasoning Decision Support: Toward a More Complete View of the Human-AI Interaction Design Space," in *CHItaly 2021: 14th Biannual Conference of the Italian SIGCHI Chapter*, Bolzano Italy, Jul. 2021, pp. 1–5.

[78] R. Dazeley, P. Vamplew, C. Foale, C. Young, S. Aryal, and F. Cruz, "Levels of explainable artificial intelligence for human-aligned conversational explanations," *Artif. Intell.*, vol. 299, p. 103525, Oct. 2021.

[79] G. Bansal *et al.*, "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*, New York, NY, USA, May 2021, pp. 1–16.

[80] M. McKenna and D. J. Coates, "Compatibilism," in *The Stanford Encyclopedia of Philosophy*, Fall 2021., E. N. Zalta, Ed. Metaphysics Research Lab, Stanford University, 2021.

[81] D. J. Simons, "Current Approaches to Change Blindness," *Vis. Cogn.*, vol. 7, no. 1–3, pp. 1–15, Jan. 2000.

[82] Z. Buçinca, P. Lin, K. Z. Gajos, and E. L. Glassman, "Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems," in *Proceedings of the 25th International Conference on Intelligent User Interfaces*, Cagliari Italy, Mar. 2020, pp. 454–464.

[83] J. Schaffer, J. O'Donovan, J. Michaelis, A. Raglin, and T. Höllerer, "I can do better than your AI: expertise and explanations," in *Proceedings of the 24th International Conference on Intelligent User Interfaces*, New York, NY, USA, Mar. 2019, pp. 240–251.

APPENDICES

Appendix 1: Lexicon of cognitive biases

Cognitive bias denomination	Definition	Ref. in the corpus
Ambiguity aversion	"The tendency to prefer known risks over unknown risks" [17]	[17]
Attention to abnormality	"People mostly ask for explanations of events that they find unusual or abnormal" [13]	[13], [34]
Attention to confidence levels	People need confidence levels to make better use of ML-assisted decision-making systems. "Prospect Theory suggests that uncertainty (or risk) is not considered independently but together with the expected outcome" [32]	[13], [32]
Attention to demographic features	Tendency to fixate on demographic features in explanations such as age and race.	[33]
Attention to foil	"Explanations are sought in response to particular counterfactual cases, which are termed foils. That is, people do not ask why event P happened, but rather why event P happened instead of some event Q." [13]	[13], [34], [35]
Attention to intentionality and responsibility	People tend to focus on intentional actions rather than non-intentional ones to select an event as a cause in a causal chain. Similarly, "an event considered more responsible for an outcome is likely to be judged as a better explanation than other causes."	[13], [34]
Attention to necessity, sufficiency and robustness	Events that are necessary, sufficient and robust to some changes are more likely to be selected as a cause.	[13]
Automation bias / overreliance	The tendency to over rely on machine's predictions.	[33], [42], [48], [54], [59], [79]
Availability bias	The tendency to believe that examples and events that easily come to mind are more representative than is actually the case.	[12], [17], [40]
Averaging bias	"Using the average of probabilities of two events for the estimation of the probability of a conjunction of the two events." [17]	[17]
Backfire effect	"Corrections of misperceptions may enhance people's false beliefs" [70]	[54]
Base-rate neglect	"The tendency to underweight evidence provided by base rates."[17]	[17]
Change blindness	"Humans inability to notice all of the changes in a presented medium." [46], [81]	[46]
Choice overload	The difficulty to make a choice when facing many choices for people of the type "mazimizer". As a consequence, they are less committed to their choices, display lower satisfaction with their choices.	[56]
Cognitive dissonance	The tendency to agree with the AI's suggestions, while being aware to have a different opinion.	[48]
Completeness bias	Longer explanations tend to lead more to overreliance than shorter ones.	[9], [41]–[43], [54]
Confirmation bias / hindsight bias	"The tendency to seek supporting evidence for one's current hypothesis." [17]	[12], [17], [42], [43], [59], [60]
Confusion of the inverse	"The mistake of confusing the confidence of an implication $A \rightarrow B$ with its inverse $B \rightarrow A$. "[17]	[17]
Conjunction fallacy	Estimating the conjunction of two statements to be more probable than one of the two statements.	[9], [17], [34]
Default bias / Status quo bias	"The tendency to favor the default option and thus the proposed suggestion" [60]	[60]
Disjunction fallacy	"Judging the probability of an event as higher than the probability of a union of the event with another event."[17]	[17]
Escalation of commitment	"People stick to a choice they made despite understanding the logical implication that doing so might lead to undesirable consequences" [60]	[60]
Familiarity bias	"Unfamiliar information might induce a reinforcement effect that causes users to avoid interacting with various content" [43]	[43]
Framing bias	People decide on options based on whether they are presented with positive or negative connotations or whether they are presented after or before the AI recommendation.	[17], [32], [68], [79]
Homunculus bias	People tend to attribute human traits to machines and therefore expect AI explanations to use the same conceptual framework used to explain human behaviors.	[13], [34]
Humans pay more attention to False Negatives to than to False Positives	"Users pay less attention to FP explanation errors and in turn, are more critical for FN explanation errors."[45]	[45]
Humans rate different-looking saliency techniques differently	Human judgment ratings of explanations are biased toward visual appearance.	[45]
Illusion of Explanatory Depth	People think they know how complex concepts work in much more depth than they actually do.	[10], [52], [59]
Illusion of validity	"Unjustified sense of confidence and hence failure when evaluating different possibilities" [61]	[61]
Inherence bias	"Humans tend to construct explanations based on accessible information about the inherent properties of a particular phenomenon instead of inaccessible information about extrinsic factors." [36]	[13], [36]
Information overload	"Providing too much information at once can result in reduced accuracy" [61]	[39], [40], [59], [61]
Insensitivity to sample size	When both confidence and support are stated, confidence scores positively affects plausibility and support is largely ignored.	[9], [17]
Insensitivity to sample variance	"Users are primarily guided by the mean and the number of ratings, and to lesser degree by the variance and origin of a rating" [56]	[56]
Mere exposure effect	The increase of trust in an AI suggestion following the mere exposure of an explanation.	[8], [17], [54]
Misunderstanding of "and"	"People interpret "and" differently than logical conjunction" [17]	[9], [17]
Misunderstanding of Boolean logic	The 'true' and 'false' conditions are perceived as hard to interpret and non-intuitive.	[40]
Misunderstanding of confidence scores	Not understanding what the confidence scores refer to.	[42]
Narration bias / Correlation vs. causation / Over-generalization	People tend to make causal narratives about everything	[40], [52], [55], [67]
Negativity bias	Users pay more attention to negative features in the AI or the AI explanations which may lead to eroding trust and pay more attention to negative outcomes.	[11], [17], [38], [40], [62]
Perceived goal impediment	"People in highly critical decision-making environments are likely to be in a serious-minded state, where additional information might be prone to being perceived as a goal impediment." [59]	[59]
Pre-use algorithmic optimism	Before using the XAI system, users had positive inferences about algorithmic capability, which disappeared after using it.	[23]
Preference for "broad" explanations	People prefer broad explanations, that explain more observations.	[13]
Preference for more complete explanations	People tend to prefer complete explanations over sound ones. Complete explanations help them form better mental models.	[41]
Preference for "simple" explanations	People prefer simple explanations to complex ones.	[13], [38]–[40]
Preference for usability vs. performance	User performance and preference on proxy tasks may not accurately predict their performance and preference on the actual decision-making tasks where their cognitive focus is elsewhere, and they can choose whether and how much to attend to the AI.	[33], [43], [82]
Primacy effect / Anchoring bias	People quickly form opinions about something based on the first information we receive about it	[11], [12], [17], [59]
Recognition bias	Recognizing information makes the user more likely to trust the explanation.	[9], [17], [35], [43]
Redundancy aversion	Redundant information is another cause of skipping explanations, making users lose trust in the explanations	[59]
Reinforcement effect / Reiteration effect	The increase of trust following repetition.	[17]
Representativeness bias	the similarity of objects or events confuses people's thinking regarding the probability of an outcome	[9], [12], [17], [40], [52]
Tendency to believe persuasive claims unsupported by evidence	Tendency to believe persuasive claims unsupported by evidence.	[48]
Unit bias	"The tendency to give a similar weight to each unit rather than weigh it according to its size." [17]	[17]
Weak evidence effect	"Weak argument in favor of a statement can lead to decreased believability of the statement." [17]	[17], [9]

Appendix 2: Classification of AI types, Explanation types, User types and Task types used in our corpus

AI types

Deep learning models	Deep reinforcement learning [55]; RoBERTa [79]; Re-ID networks [36] ; BERT [54]; CNN VGG-19 [45]; deep neural network based on GoogleNet and cutset network [11]
Shallow models	LASSO regression [40]; GAM / sLM [39]; Decision trees, logistic regression, shallow (1- to 2-layer) neural network [62]; Random forest classifier [10], [48]; GAM and gradient boosted decision trees (LightGBM) [52]; SVM [33], [54] ; linear regression [23] ;Multi-label gradient boosted tree [12]; k-nearest neighbor and baged decision tree [41];
Wizard of Oz	[42], [44], [69]

Explanation types

Local feature importance	Saliency map [55], [72]; word highlighting [23], [33], [54]; Other input-based interface [8], [11], [35], [83]; roBERTa+LIME [79]; sensitivity analysis MOEA/D [12] ; SHAP [10], [12], [52]; COGAM: "simplified" GAMs and sparse LM [39]; not specified [43]; list of contributing features [40], [41], [69], [82]; GAM[52]; comprehensive and selective input list [42]
Rule-based	not specified [17]; Apriori algorithm and top-down greedy hill-climbing algorithm [9]; manual deductive explanation [82]
Example-based	social proof [35]; collaborative explanation[56]; manual inductive explanation [82]; not specified [40], [59]; MMD-critic [54]; nearest neighbours [41]
Counterfactuals	LORE [12]; interactive counterfactual [33], [40] ; not specified [35], [59]
Natural language explanations	expert-generated explanations [79]; automatic text-based justifications [36], [48], [60]
Confidence estimates	global [54], [68]; local [41], [42], [44], [79]; not specified [59]
Global explanations	distribution of values [40]; decision tree [41], [62]; output visualization [62]

User types

Domain expert	[12], [40], [42], [59]–[61]
Machine Learning Expert	[43], [52]
Lay user	[9]–[11], [23], [33], [35], [35], [39], [41], [43], [45], [48], [54], [56], [61], [68], [69], [79], [82], [83]
Researcher	[55]

Tasks and domains

Artificial	Sentiment analysis of book and beer reviews [79]; Prediction of fat content in a food image [44], [82]; Prediction of traffic accidents in a country, Prediction of quality of living in a city, Movie rating, Mushroom poisonous/edible prediction [9]; The Desert Survival Problem [68]; The Diner's Dilemma game [83]
Law & Regulation	child welfare screening [40]; Identity recognition [36]; recidivism prediction [33]
Business & Finance	House price estimate [39]; credit scoring [10], [52]
Education	LSAT question answering [79]
Leisure	chess playing [60]; Hotel rating [56]; Reading time prediction [43], music recommender [41]
Healthcare	Medical diagnosis [12], [62]; Prediction of balance disorders [42]; Recommendation of a medical prescription [59], [69]; Symptom checkers [35]
Others	fake news detection [48]; Application to lose weight [8]; Deception detection in hotel reviews [54]; profession prediction [33]; Image recognition [45]; activity recognition in video (in cooking videos) [11]; emotional analysis [23]