International Journal of Academic Information Systems Research (IJAISR)
ISSN: 2643-9026
Vol. 5 Issue 3, March - 2021, Pages: 8-19

A Review Of Possible Effects Of Cognitive Biases On Interpretation Of Rule-Based Machine Learning Models.

Sherzod Yarashev[1*], Elyor G'aybulloyev[1], Begzod Erkinov[1], Temur Ochilov[1], Jurabek Abdiyev[2]

[1]Tashkent University of Information Technologies named after Muhammad al-Khwarizmi, Uzbekistan, Tashkent, Amir Temur street 108.

[2]Physical-technical Institute of NPO "Physics – Sun" of Uzbekistan Academy of Sciences Uzbekistan, Tashkent, Chingiz Aitmatov street 2B.

Author: sherzodyarashev1997@gmail.com (Sh. Yarashev)

Abstract: *While the interpretability of machine learning models is often equated with their mere syntactic comprehensibility, we think that interpretability goes beyond that, and that human interpretability should also be investigated from the point of view of cognitive science. The goal of this paper is to discuss to what extent cognitive biases may affect human understanding of interpretable machine learning models, in particular of logical rules discovered from data. Twenty cognitive biases are covered, as are possible debiasing techniques that can be adopted by designers of machine learning algorithms and software. Our review transfers results obtained in cognitive psychology to the domain of machine learning, aiming to bridge the current gap between these two areas. It needs to be followed by empirical studies specifically focused on the machine learning domain.*

Keywords – Cognitive bias, Cognitive illusion, Interpretability, Machine learning, Rule induction.

1. Introduction.

This paper aims to investigate the possible effects of cognitive biases on human understanding of machine learning models, in particular inductively learned rules. We use the term "cognitive bias" as a representative for various cognitive phenomena that materialize themselves in the form of occasionally irrational reasoning patterns, which are thought to allow humans to make fast judgments and decisions.

Their cumulative effect on human reasoning should not be underestimated as "cognitive biases seem reliable, systematic, and difficult to eliminate" [83]. The effect of some cognitive biases is more pronounced when people do not have well-articulated preferences [168], which is often the case in explorative data analysis.

Previous works have analyzed the impact of cognitive biases on multiple types of human behavior and decision making. A specific example is the seminal book "Social cognition" by Kunda [90], which is concerned with the impact of cognitive biases on social interaction. Another, more recent work by Serfas [147] focused on the context of capital investment. Closer to the domain of machine learning, in their article "Psychology of Prediction", Kahneman and Tversky [84] warned that cog-nitive biases can lead to violations of the Bayes theorem when people make fact-based predictions under uncertainty. These results directly relate to inductively learned rules, since these are associated with measures such as confidence and support expressing the (un)certainty of the prediction they make. Despite some early work [104,105] showing the importance of study of cognitive phenomena for rule induction and machine learning in general, there has been a paucity of follow-up research. In previous work [53], we have evaluated a selection of cognitive biases in the very specific context of whether minimizing the complexity or length of a rule will also lead to increased interpretability, which is often taken for granted in machine learning research.

In this paper, we attempt to systematically relate cognitive biases to the interpretation of machine learning results. We anchor our discussion on inductively learned rules, but note in passing that a deeper understanding of human cognitive bi-ases is important for all areas of combined human-machine decision making. We focus primarily on symbolic rules because they are generally considered to belong to the class of interpretable models, so that there is little general awareness that different ways of presenting or formulating them may have an important impact on the perceived trustworthiness, safety, or fairness of an AI system. In principle, our discussion also applies

to rules that have been inferred by deduction, where, however, such concerns are maybe somewhat alleviated by the proved correctness of the resulting rules. To further our goal, we review twenty cognitive biases and judgmental heuristics whose misapplication can lead to biases that can distort the interpretation of inductively learned rules. The review is intended to help to answer questions such as: *How do cognitive biases affect the human understanding of symbolic machine learning models? What could help as a "debiasing antidote"?*

This paper is organized as follows. Section 2 provides a brief review of related work published at the intersection of rule learning and psychology. Section 3 motivates our study by showing an example of a learnt rule and discussing sample cognitive biases that can affect its plausibility. Section 4 describes the criteria that we applied to select a subset of cognitive biases into our review, which eventually resulted in twenty biases. These biases and their respective effects and causes are covered in detail in Section 5. Section 6 provides a concise set of recommendations aimed at developers of rule learning algorithms and user interfaces. In Section 7 we state the limitations of our review and outline directions for future work. The conclusions summarize the contributions of the paper.

2. Background and related work.

We selected individual rules as learnt by many machine learning algorithms as the object of our study. Focusing on simple artefacts—individual rules—as opposed to entire models such as rule sets or rule lists allows a deeper, more focused analysis since a rule is a small self-contained item of knowledge. Making a small change in one rule, such as adding a new condition, allows to test the effect of an individual factor. In this section, we first motivate our work by putting it into the context of prior research on related topics. Then, we proceed by a brief introduction to inductive rule learning (Section 2.2) and a brief recapitulation of previous work in cognitive science on the subject of decision rules (Section 2.3). Finally, we introduce cognitive biases (Section 2.4) and rule plausibility (Section 2.5), which is a measure of rule comprehension.

2.1. Motivation.

In the following three paragraphs, we discuss our motivation for this review, and summarize why we think this work is relevant to the larger artificial intelligence community.

Rules as interpretable models Given that neural networks and ensembles of decision trees are increasingly becoming the prevalent type of representation used in machine learning, it might be at first surprising that our review focuses almost exclusively on decision rules. The reason is that rules are widely used as a means for communicating explanations of a variety of machine learning approaches. In fact, quite some work has been devoted to explaining black-box models, such as neural networks, support vector machines and tree ensembles with interpretable surrogate models, such as rules and decision trees (for a survey on this line of work we refer, e.g., to [69]). As such a conversion typically also goes hand-in-hand with a corresponding reduction in the accuracy of the model, this approach has also been criticized [142], and the interest in directly learning rule-based models has recently renewed (see, e.g., [52,176,110,173]).

Embedding cognitive biases to learning algorithms The *applications* of cognitive biases go beyond explaining existing machine learning models. For example, Taniguchi et al. [159] demonstrate how a cognitive bias can be embedded in a machine learn-ing algorithm, achieving superior performance on small datasets compared to commonly used machine learning algorithms with "generic" inductive bias.

Paucity of research on cognitive biases in artificial intelligence Several recent position and review papers on explainability in Artificial Intelligence (xAI) recognize that cognitive biases play an important role in explainability research [106,126]. To our knowledge, the only systematic treatment of psychological phenomena applicable to machine learning is provided by the review of Miller [106], which focuses on reasons and thought processes that people apply during explanation selection, such as causality, abnormality and the use of counterfactuals. This authoritative review observes that there are currently no studies that look at cognitive biases in the context of

selecting explanations. Because of the paucity of applicable research focusing on machine learning, the review of Miller [106]—like the present paper—takes the first step of applying influential psychological studies to explanation in the xAI context without accompanying experimental validation specific to machine learning. While Miller [106] summarizes the main reasoning processes that drive generation and understanding of explana-tions, our review focuses specifically on cognitive biases as psychological phenomena that can distort the interpretation of machine learning models if not properly accounted for. The role of bias mitigation in machine learning has been recently recognized in [175], who describe four biases applicable to machine learning and for each propose a specific debiasing strategy.

IF A AND B THEN C
confidence=c and support=s
IF veil is white AND odor is foul THEN mushroom is poisonous
confidence = 90%, support = 5%

Fig. 1. Inductively learned rule.

Our review is more comprehensive as we include twenty biases and we also provide a more detailed analysis of each of the biases included.

2.2. Decision rules in machine learning.

An example of an inductively learned decision rule, which is a subject of the presented review, is shown in Fig. 1. Following the terminology of Fürnkranz et al. [52], *A*, *B*, *C* represent *literals*, i.e., Boolean expressions which are composed of attribute name (e.g., veil) and its value (e.g., white). The conjunction of literals on the left side of the rule is called *antecedent* or *rule body*, the single literal predicted by the rule is called *consequent* or *rule head*. Literals in the body are sometimes referred to as *conditions* throughout the text, and the consequent as the *target*. While this rule definition is restricted to conjunctive rules, other definitions, e.g., the formal definition given by Slowinski et al. [152], also allow for negation and disjunction as connectives.

Rules in the output of rule learning algorithms are most commonly characterized by two parameters, confidence and support. The *confidence* of a rule—sometimes also referred to as *precision*—is defined as $a/(a + b)$, where a is the number objects that match both the conditions of the rule as well as the consequent, and b is the number of objects that match the antecedent but not the consequent. The *support* of a rule is either defined as a/N, where N is the number of all objects (relative support), or simply as a (absolute support). A related measure is *coverage*, which is the total number of objects that satisfy the body of the rule ($a + b$).

The values of support and confidence are often used as indications of how subjectively interesting the given rules is. Research has shown the utility of involving thresholds on a range of additional measures of significance [121]. Out of the dozens of proposed formulas, the one most frequently adopted seems to be the *lift* measure, which is a ratio of the confidence of the rule and the probability of occurrence of the head of the rule (not considering the body). If lift is greater than 1, this indicates that the rule body and the rule head appear more often together than would correspond to chance.

In the special case of learning rules for the purpose of building a classifier, the consequent of a rule consists only of a single literal, the so-called *class*. In this case, a is also known as the number of *true positives*, and b as the number of *false positives*.

Some rule learning frameworks, in particular association rule learning [1,188], require the user to set thresholds for minimum confidence and support. Only rules with confidence and support values meeting or exceeding these thresholds are included on the output of rule learning and presented to the user.

Even though the terminology, “support” and “confidence”, is peculiar to symbolic rule learning (in particular to asso-ciation rule mining), the underlying concepts are universally adopted. For example, they are essentially equivalent to the terms “recall” and “precision” commonly used in information retrieval and correspond to the concepts of“accuracy” and “coverage” of general machine learning models.

2.3. Decision rules in cognitive science.

Rules are used in commonly embraced models of human reasoning in cognitive science [153,118,130]. They also closely relate to Bayesian inference, which also frequently occurs in models of human reasoning. Consider the first rule of Fig. 1. This rule can be interpreted as a hypothesis corresponding to the logical implication $A \wedge B \rightarrow C$. We can express the plausibility of such a hypothesis in terms of Bayesian inference as the conditional probability $\Pr(C \mid A, B)$. This corresponds to the confidence of the rule, as used in machine learning and as defined above, and to the *strength of evidence*, a term used by cognitive scientists [165].

Given that $\Pr(C \mid A, B)$ is a probability estimate computed on a sample, another relevant piece of information for deter-mining the plausibility of the hypothesis is the robustness of this estimate. This corresponds to the number of instances for which the rule has been observed to be true. The size of the sample (typically expressed as a ratio) is known as rule support in machine learning and as the *weight of the evidence* in cognitive science [165].[1]

Psychological research on hypothesis testing in rule discovery tasks has been performed in cognitive science at least since the 1960s. The seminal article by Wason [177] introduced what is widely referred to as *Wason's 2-4-6* task. Participants are

1 Interestingly, balancing the likelihood of the judgment and the weight of the evidence in the assessed likelihood was already studied by Keynes [86] (according to Camerer and Weber [22]). given the sequence of numbers 2, 4 and 6 and asked to find out the rule that generated this sequence. In the search for the hypothesized rule, they provide the experimenter other sequences of numbers and the experimenter answers whether the provided sequence conforms to the rule, or not. While the target rule is simple "ascending sequence", people find it difficult to discover this specific rule, presumably because they use the *positive test strategy*, a strategy of testing a hypothesis by examining evidence confirming the hypothesis at hand rather than searching for disconfirming evidence [87]. For example, if they have the hypothesis that the rule is a sequence of numbers increasing by two, they can provide a sequence 3-5-7, trying to confirm the hypothesis, rather than a sequence, such as 1-2-3, looking for an alternative hypothesis.

2.4. Cognitive bias.

According to the Encyclopedia of Human Behavior [181], the term cognitive bias was introduced in the 1970s by Amos Tversky and Daniel Kahneman [165], and is defined as a "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."

The narrow initial definition of cognitive bias as a shortcoming of human judgment was criticized by German psychologist Gerd Gigerenzer, who started in the late 1990s the "Fast and frugal heuristic" program to emphasize ecological rationality (validity) of judgmental heuristics [62]. According to this research program, cognitive biases often result from an application of a heuristic in an environment for which it is not suited rather than from problems with heuristics themselves, which work well in usual contexts.

In the present view, we define cognitive biases and associated phenomena broadly. We include cognitive biases related to thinking, judgment, and memory. We also include descriptions of thinking strategies and judgmental heuristics that may result in cognitive biases, even if they are not necessarily biases themselves.

Debiasing An important aspect related to the study of cognitive biases is the validation of strategies for mitigating their effects in cases when they lead to incorrect judgment. A number of such *debiasing* techniques have been developed, with researchers focusing intensely on the clinical and judicial domains (cf. e.g. [93,27,99]), apparently due to costs associated with erroneous judgment in these fields. Nevertheless, general debiasing techniques can often be derived from such studies.

The choice of an appropriate debiasing technique typically depends on the type of error induced by the bias, since this implies an appropriate debiasing strategy [5]. Larrick [92] recognizes the following three categories: psychophysically-based error, association-based error, and strategy-based error. The first two are attributable to the unconscious, automatic processes, sometimes referred to as "System 1". The last one is attributed to reasoning

processes (System 2) [38]. For biases attributable to System 1, the most generic debiasing strategy is to shift processing to the conscious System2 [96], [148, p. 491].

Another perspective on debiasing is provided by Croskerry et al. [27], who organize debiasing techniques by their way of functioning, rather than the bias they address, into the following three categories: educational strategies, workplace strategies and forcing functions. While Croskerry et al. [27] focused on clinicians, our review of debiasing aims to be used as a starting point for analogous guidelines for an audience of machine learning practitioners. For example, the general workplace strategies applicable in the machine learning context include group decision making, personal accountability, and planning time-out sessions to help slowing down. All of these strategies could lead to a higher probability of activating System 2 and thus reducing the biases which originate in the failure of System 1.

Function and validity of cognitive biases The function of cognitive biases is a subject of scientific debate. According to the review of functional views by Pohl [131], there are three fundamental positions among researchers. The first group considers them as dysfunctional errors of the system, the second group as faulty by-products of otherwise functional processes, and the third group as adaptive and thus functional responses. According to Pohl [131], most researchers are in the second group, where cognitive biases are considered to be "built-in errors of the human information-processing systems".

In this work, we consider judgmental heuristics and cognitive biases as strategies that evolved to improve the fitness and chances of survival of the individual in particular situations or as consequences of such strategies. This defense of biases is succinctly expressed by Haselton and Nettle [71]: "Both the content and direction of biases can be predicted theoretically and explained by optimality when viewed through the long lens of evolutionary theory. Thus, the human mind shows good design, although it is designed for fitness maximization, not truth preservation."

According to the same paper, empirical evidence shows that cognitive biases are triggered or strengthened by environ-mental cues and context [71]. Given that the interpretation of machine learning results is a task unlike the simple automatic cognitive processes to which a human mind is adapted, cognitive biases are likely to have an influence upon it.

2.5. Measures of interpretability, perceived and objective plausibility.

We claim that cognitive biases can affect the interpretation of rule-based models. However, how does one measure interpretability? According to our literature review, there is no generally accepted measure of interpretability of machine learning models. Model size, which was used in several studies, has recently been criticized [48,156,53] primarily on the grounds that the model's syntactic size does not capture any aspect of the model's semantics. A particular problem related to semantics is the compliance to pre-existing expert knowledge, such as domain-specific monotonicity constraints.

In prior work [53], we embrace the concept of *plausibility* to measure interpretability. In the following, we will briefly introduce this concept because in the remainder of this article, we will use some material collected[2] in user studies reported on in [53] to illustrate some of the discussed biases and cognitive phenomena. The word 'plausible' is defined according to the Oxford Dictionary of US English as "seeming reasonable or probable" and according to the Cambridge dictionary of UK English as "seeming likely to be true, or able to be believed". We can link the inductively learned rule to the concept of "hypothesis" used in cognitive science. There is a body of work in cognitive science on analyzing the perceived plausibility of hypotheses [58,59,4].

In a recent review of interpretability definitions by Bibal and Frénay [17], the term plausibility is not explicitly covered, but a closely related concept of *justifiability* is stated to depend on interpretability. Martens et al. [98] define justifiability as "intuitively correct and in accordance with domain knowledge". By adopting plausibility, we address the concern expressed in Freitas [48] regarding the need to reflect domain semantics when interpretability is measured.

3. Motivational example.

When an analyst[3] evaluates the plausibility of a rule, a number of biases can be triggered by different facets of the rule. Consider, e.g., Fig. 2 which shows how the interpretation of a rule predicting whether a movie will get a good rating based on its release date, genre, and director can be affected by various cognitive biases. The analyst needs to evaluate whether each attribute and value is predictive for the target (good movie rating) and how large set of movies it delimits. Additionally, the analyst needs to correctly process the syntactical elements in the rule (here AND and THEN), realizing that AND acts as a set intersection. Finally, the analyst needs to understand the confidence and support values and their trade-offs. Fig. 2 shows several illustrative cognitive biases for each of these processes. Two of them are discussed in greater detailed below, all of these (as well as many other) are covered in much greater detail in Section 5.

Information bias According to *information bias*, more information can make a rule look more plausible even if this informa-tion is irrelevant. In this case, the analyst may not know the director John Smith or any of his movies, but nevertheless, the rule that includes this condition may appear to be more plausible to the analyst than the same rule without this condition just because it involves more information.

Insensitivity to sample size According to the *insensitivity to sample size effect* [165] there is a systematic bias in human thinking that makes humans overestimate the strength of evidence (confidence) and underestimate the weight of evidence (support). The example rule in Fig. 2 is associated with values of confidence and support that inform about the strength and weight of evidence. While it seems to have an excitingly high confidence (100%), its low support indicates that this high value may be deceptive, as is sometimes the case for rule learning algorithms [7]. This crucially depends on the absolute and not the relative support of this rule: if this rule originates from a database with a million of movies, a support of 1% corresponds to 10,000 movies, whereas the same rule may only be based on a single movie in a database with 100 entries. Yet, the low support may be largely ignored by the analyst due to insensitivity to sample size.

Opposing effects of biases Sometimes the same piece of information can trigger opposing biases. Communicating the iden-tity of the director of the movie can increase the plausibility due to the information bias, but if the specific director is not known to the user also decrease it through the ambiguity aversion bias. Fig. 2 also refers to the primacy effect and misunderstanding of "and", where the impact on plausibility largely depends on the context. For example, the year of the movie as the first provided information is overweighted due to the primacy effect, but the overall effect on plausibility will depend on how this information is perceived and aggregated by the analyst, which is determined by other factors.

Debiasing Whether the biases listed in Fig. 2 apply depends, among other factors, on the analyst's background knowledge, quality of reasoning skills and statistical sophistication. An analysis of relevant literature from cognitive science not only reveals applicable biases but also sometimes provides methods for removing or limiting their effect (debiasing). Several debiasing techniques are also illustrated in Fig. 2. In principle, we found three categories of debiasing techniques: 1. Training users, 2. Adapting learning algorithms, 3. Adapting the representation of the model and/or the user interface.

1.Training users In [47] (cf. also [119,138]) it was shown that training can significantly improve statistical reasoning and help people better understand the importance of sample size ('law of large numbers'), which is instrumental for correctly interpreting statistical properties such as rule support and rule confidence.

[2] In [53], we present quantitative results for several selected biases, whereas the current article presents much broader review of the available literature. For illustrating some of the claims, we also make use of some of the textual responses from the participants, which were not featured in the above-mentioned work.

[3] The prospective human users of rule models are often called "analysts" in this article.

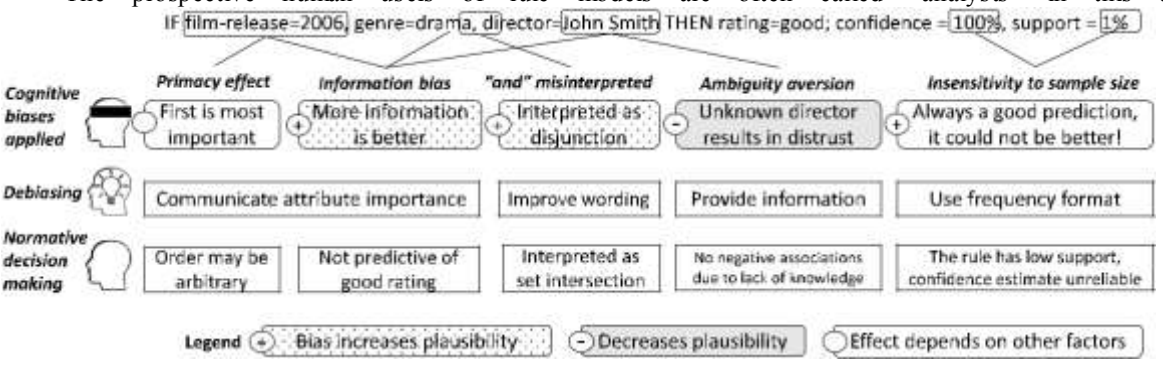

Fig. 2. Cognitive biases affecting perceived plausibility of example learnt rule.

Mined rule:

IF film-release=2006 AND genre=drama
AND director="John Smith" THEN rating=good
confidence = 100%, support = 1%

Verbalized rule:

If a film is released in *2006* **and also** Genre is *drama* **and also** the director is John Smith then its rating is *good*.

Improved explanation of rule:

In our data, there are *2 movies* which match the conditions of this rule. Out of these, *2* are correctly classified as being *good*. The rule thus makes the correct prediction in *2/2 = 100*% percent of cases, which corresponds to the confidence of the rule. The complete database contains *200 movies*, out of these, the current rule correctly classifies 2. The support of the rule is thus *2/200 = 1*%

Fig. 3. Suggested general frequency-based representation of an association rule.

2.Adapting learning algorithms One possibility in terms of adaptation of learning algorithms is to compute confidence inter-vals for rule confidence as proposed, e.g., in [179]. The support of a rule would then be—in a way—directly embedded into the presentation of rule confidence [102]. Spurious rules with little statistical grounding may not be shown to the user at all.

3. *Adapting the representation* A common way used in rule learning software for displaying rule confidence and support metrics is to use percentages, as in our example. Extensive research in psychology has shown that if frequencies are used instead, then the number of errors in judgment drops [61,63]. Reflecting these suggestions, the hypothetical rule learnt from our movies recommendation dataset could be presented as shown in Fig. 3.

Rules can be presented in different ways (as shown), and depending on the way the information is presented, humans may perceive their plausibility differently. In this particular example, confidence is no longer conveyed only as a percentage "100%" but also using the expression "2 out of 2". Support is presented as an absolute number (2) rather than just a percentage (1%).

A correct understanding of machine learning models can be difficult even for experts. In this section, we tried to demon-strate why addressing cognitive biases can play an important role in making the results of inductive rule learning more understandable. However, it should only serve as a motivational example, rather than a general guideline. In the remainder of this paper, the biases applied to our example will be revisited in greater depth, along with many other biases, and more concrete recommendations will be given.

4. Scope of survey.

A number of cognitive biases have been discovered, experimentally studied, and extensively described in the literature. As Pohl [131] states in a recent authoritative book on cognitive illusions: "There is a plethora of phenomena showing that we deviate in our thinking, judgment and memory from some objective and arguably correct standard." This book covers 24 cognitive biases, and even 51 biases are covered by Evans [37].

We first selected a subset of biases which would be reviewed. To select applicable biases, we looked for those that can interact with the following properties of rules, and their activation could result in an impact on perceived plausibility of rules: 1. rule length (the number of literals in an antecedent), 2. rule interest measures (especially support and confidence),
3. order of conditions in a rule and order of rules in the rule list, 4. specificity and predictive power of conditions (correlation with a target variable), 5. use of additional logical connectives (conjunction, disjunction, negation), 6. treatment of missing information (inclusion of conditions referring to missing values), and 7. conflict between rules in the rule list.

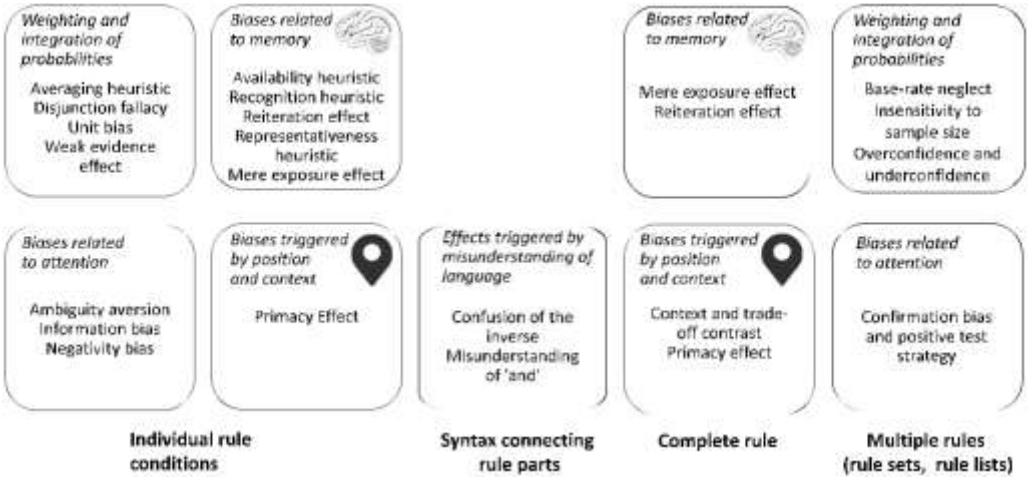

Fig. 4. Cognitive biases grouped by the affected or triggering rule model element and by their underlying mechanism.

Through a selection of appropriate learning heuristics, the rule learning algorithm can influence these properties. For example, most heuristics implement some form of a trade-off between the coverage or support of a rule, and its implication strength or confidence [51,52].

While doing the initial selection of cognitive biases to study, we tried to identify those most relevant for machine learning research matching our criteria. In the end, our review focused on a selection of 20 judgmental heuristics and cognitive biases. Future work might focus on expanding the review with additional relevant biases, such as labeling and overshadowing effects [131].

5. Review of cognitive biases.

In this section, we cover a selection of twenty cognitive biases. For all of them, we include a short description including an example of a study demonstrating the bias and its proposed explanation. We pay particular attention to their potential effect on the interpretability of rule learning results, which has not been covered in previous works. Fig. 4 shows a high-level overview of the results of our analysis. The figure organizes the surveyed biases according to the primary affected element of rule models, ranging from conditions (literals) as the basic building block, to entire rules and rule models. The second perspective conveyed in the figure relates to the underlying mechanism of the biases.

In a recent scientometric survey of research on cognitive biases in information systems [45], no articles are mentioned that aim at machine learning. For general information systems research, the authors claim that "most

articles' research goal [is] to provide an explanation of the cognitive bias phenomenon rather than to develop ways and strategies for its avoidance or targeted use". In contrast, our review aims at the advancement of the field beyond the explanation of applicable phenomena, by also discussing specific debiasing techniques.

For all cognitive biases, we thus suggest a debiasing technique that could be effective in aligning the perceived plausibil-ity of the rule with its objective plausibility. While we include a description only of the most prominent debiasing strategies for each bias, it is possible that some of the debiasing strategies may be more general and could be effective for multiple biases.

The utility of this article for the reader would increase if concrete proposals for debiasing machine learning (rule learn-ing) results were included. However, there is a paucity of applicable work on debiasing techniques applied specifically to machine learning (or directly on rule learning), and the invention of a rule-learning-specific debiasing technique for each of the twenty surveyed biases is out of the scope of this article. We, therefore, decided to introduce only one debiasing technique developed specifically for rule learning, choosing an approach that is to our knowledge the most well studied.

We chose the "frequency format" debiasing method, which has been documented to reduce the number of bias-induced judgment errors across a variety of different tasks as supported by a body of psychological studies. We adapted this method for rule learning and used in a user study reported in [53]. Since then, it has been subject of at least one other user study focused specifically on debiasing methods for rule learning. This method is introduced already in Section 5.1 in the context of the motivating Linda problem and the representativeness heuristic. However, since it provides a way for expressing rule confidence and rule support, it is primarily intended as a debiasing method for the base rate neglect (Section 5.5) and the insensitivity to sample size (Section 5.6). The proposed generalized adaptation to rule learning also adopts recommendations from psychological research for addressing the misunderstanding of *'and'* (Section 5.2).

An overview of the main features of the reviewed cognitive biases is presented in Table 1. Note that the debiasing techniques that we describe have only limited grounding in applied psychological research and require further validation,

Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.

Which is more probable?

(a) Linda is a bank teller.

(b) Linda is a bank teller and is active in the feminist movement.

Fig. 5. Linda problem.

since as Lilienfeld et al. [96] observe, there is a general paucity of research on debiasing in psychological literature, and the existing techniques suffer from a lack of theoretical coherence and mixed research evidence concerning their efficacy.

5.1. Conjunction fallacy and representativeness heuristic.

The conjunction fallacy refers to a judgment that is inconsistent with *the conjunction rule* – the probability of conjunction, Pr*(A, B)*, cannot exceed the probability of its constituents, Pr*(A)* and Pr*(B)*. It is often illustrated with the "Linda" problem in the literature [167]. In the Linda problem, depicted in Fig. 5, subjects are asked to compare conditional probabilities Pr*(B | L)* and Pr*(F , B | L)*, where *B* refers to "bank teller", *F* to "active in feminist movement" and *L* to the description of Linda [9].

Multiple studies have shown that people tend to consistently select the second hypothesis as more probable, which is in conflict with the conjunction rule. In other words, it always holds for the Linda problem that

$\Pr(F, B \mid L) \leq \Pr(B \mid L)$.

Fig. 6. Augmented screenshot from the EasyMiner rule learning system. A. User first sets a template to which the discovered rules must comply. Rules must also match the set minimum support and confidence thresholds; B. List of discovered rules; C. User chooses one rule; D. contingency table for chosen rule; E. (highlighted text) Proposed addition to the user interface – inclusion of class priors. Note that while in the rest of the article a literal is denoted as, e.g., cap_shape=k, the EasyMiner system represents it as cap_shape(k) following the notation of the GUHA method [136].

A closely related phenomenon is the *positive test strategy* (PTS) described by Klayman and Ha [87]. This reasoning strategy suggests that when trying to test a specific hypothesis, people examine cases which they expect to confirm the hypothesis rather than the cases which have the best chance of falsifying it. The difference between PTS and confirmation bias is that PTS is applied to test a candidate hypothesis while confirmation bias is concerned with hypotheses that are already established [123, p. 93]. The experimental results of Klayman and Ha [87] show that under realistic conditions, PTS can be a very good heuristic for determining whether a hypothesis is true or false, but it can also lead to systematic errors if applied to an inappropriate task.

Implications for rule learning This bias can have a significant impact depending on the purpose for which the rule learning results are used. If the analyst has some prior hypothesis before obtaining the rule learning results, according to the confir-mation bias the analyst will tend to "cherry pick" rules confirming this prior hypothesis and disregard rules that contradict it. Given that some rule learners may output contradicting rules, the analyst may tend to select only the rules conforming to the hypothesis, disregarding applicable rules with the opposite conclusion, which could otherwise turn out to be more relevant.

Debiasing techniques Delaying *final* judgment and slowing down work has been found to decrease confirmation bias in several studies [154,128]. User interfaces for rule learning should thus give the user not only the opportunity to save or mark interesting rules, but also allow the user to review and edit the model at a later point in time. An example rule learning systemwith this specific functionality is EasyMiner [173].

Wolfe and Britt [186] successfully experimented with providing subjects with explicit guidelines for considering evidence both for and against a hypothesis. Provision of "balanced" instructions to search evidence for and against a given hypothesis reduced the incidence of confirmation bias from 50% exhibited by the control group to a significantly lower 27.5%. The assumption that educating users about cognitive illusions can be an effective debiasing technique for positive test strategy has been empirically validated on a cohort of adolescents by Barberia et al. [11].

References

[1] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, A.I. Verkamo, Fast discovery of association rules, in: U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy (Eds.), Advances in Knowledge Discovery and Data Mining, AAAI Press, 1995, pp. 307–328.

[2] D. Albarracín, A.L. Mitchell, The role of defensive confidence in preference for proattitudinal information: how believing that one is strong can sometimes be a defensive weakness, Pers. Soc. Psychol. Bull. 30 (2004) 1565–1584.

[3] J. Alcala-Fdez, R. Alcala, F. Herrera, A fuzzy association rule-based classification model for high-dimensional problems with genetic rule selection and lateral tuning, IEEE Trans. Fuzzy Syst. 19 (2011) 857–872.

[4] J. Anderson, D. Fleming, Analytical procedures decision aids for generating explanations: current state of theoretical development and implications of their use, J. Account. Tax. 8 (2016) 51.

[5] H.R. Arkes, Costs and benefits of judgment errors: implications for debiasing, Psychol. Bull. 110 (1991) 486.

[6] H.R. Arkes, C. Christensen, C. Lai, C. Blumer, Two methods of reducing overconfidence, Organ. Behav. Hum. Decis. Process. 39 (1987) 133–144.

[7] P.J. Azevedo, A.M. Jorge, Comparing rule measures for predictive association rules, in: Proceedings of the 18th European Conference on Machine Learning, ECML-07, Springer, Warsawa, Poland, 2007, pp. 510–517.

[8] M. Bar-Hillel, The role of sample size in sample evaluation, Organ. Behav. Hum. Perform. 24 (1979) 245–257.

[9] M. Bar-Hillel, Commentary on Wolford, Taylor, and Beck: the conjunction fallacy?, Mem. Cogn. 19 (1991) 412–414.

[10] M. Bar-Hillel, E. Neter, How alike is it versus how likely is it: a disjunction fallacy in probability judgments, J. Pers. Soc. Psychol. 65 (1993) 1119.

[11] I. Barberia, F. Blanco, C.P. Cubillas, H. Matute, Implementation and assessment of an intervention to debias adolescents against causal illusions, PLoS ONE 8 (2013) e71303.

[12] A.K. Barbey, S.A. Sloman, Base-rate respect: from ecological rationality to dual processes, Behav. Brain Sci. 30 (2007) 241–254.

[13] J. Baron, J. Beattie, J.C. Hershey, Heuristics and biases in diagnostic reasoning: II congruence, information, and certainty, Organ. Behav. Hum. Decis. Process. 42 (1988) 88–110.

[14] C.P. Beaman, R. McCloy, P.T. Smith, When does ignorance make us smart? Additional factors guiding heuristic inference, in: Proceedings of the Cognitive Science Society, 2006, pp. 54–58.

[15] E.S. Becker, M. Rinck, Reversing the mere exposure effect in spider fearfuls: preliminary evidence of sensitization, Biol. Psychol. 121 (2016) 153–159.

[16] P. Berka, Comprehensive concept description based on association rules: a meta-learning approach, Intell. Data Anal. 22 (2018) 325–344.

[17] A. Bibal, B. Frénay, Interpretability of machine learning models and representations: an introduction, in: Proceedings of the 24th European Symposium on Artificial Neural Networks, ESANN, 2016, pp. 77–82.

[18] L.E. Boehm, The validity effect: a search for mediating variables, Pers. Soc. Psychol. Bull. 20 (1994) 285–293.

[19] S.D. Bond, K.A. Carlson, M.G. Meloy, J.E. Russo, R.J. Tanner, Information distortion in the evaluation of a single option, Organ. Behav. Hum. Decis. Process. 102 (2007) 240–254.

[20] R.F. Bornstein, Exposure and affect: overview and meta-analysis of research, 1968–1987, Psychol. Bull. 106 (1989) 265.

[21] P.D. Bruza, Z. Wang, J.R. Busemeyer, Quantum cognition: a new theoretical approach to psychology, Trends Cogn. Sci. 19 (2015) 383–393.

[22] C. Camerer, M. Weber, Recent developments in modeling preferences: uncertainty and ambiguity, J. Risk Uncertain. 5 (1992) 325–370.

[23] A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E.R. Hruschka Jr., T.M. Mitchell, Toward an architecture for never-ending language learning, in: Proceedings of the 24th AAAI Conference on Artificial Intelligence, Atlanta, Atlanta, Georgia, 2010, p. 3.

[24] G. Charness, E. Karni, D. Levin, On the conjunction fallacy in probability judgment: new experimental evidence regarding Linda, Games Econ. Behav. 68 (2010) 551–556.
[25] R.T. Clemen, K.C. Lichtendahl, Debiasing expert overconfidence: a Bayesian calibration model, in: Sixth International Conference on Probabilistic Safety Assessment and Management, PSAM6, 2002.
[26] W.G. Cochran, Sampling Techniques, John Wiley & Sons, 2007.
[27] P. Croskerry, G. Singhal, S. Mamede, Cognitive debiasing 2: impediments to and strategies for change, BMJ Qual. Saf. bmjqs–2012 (2013).
[28] P.B. De Laat, Algorithmic decision-making based on machine learning from big data: can transparency restore accountability?, Philos. Technol. (2017) 1–17.
[29] A. Dechêne, C. Stahl, J. Hansen, M. Wänke, The truth about the truth: a meta-analytic review of the truth effect, Personal. Soc. Psychol. Rev. 14 (2010) 238–257.
[30] R. Deutsch, R. Kordts-Freudinger, B. Gawronski, F. Strack, Fast and fragile: a new look at the automaticity of negation processing, Exp. Psychol. 56 (2009) 434.
[31] C. Díaz, C. Batanero, J.M. Contreras, Teaching independence and conditional probability, Bol. Estad. Investig. Oper. 26 (2010) 149–162.
[32] S. Donovan, S. Epstein, The difficulty of the Linda conjunction problem can be attributed to its simultaneous concrete and unnatural representation, and not to conversational implicature, J. Exp. Soc. Psychol. 33 (1997) 1–20.
[33] U.K. Ecker, J.L. Hogan, S. Lewandowsky, Reminders and repetition of misinformation: helping or hindering its retraction?, J. Appl. Res. Mem. Cogn. 6 (2017) 185–192.
[34] S.E. Edgell, J. Harbison, W.P. Neace, I.D. Nahinsky, A.S. Lajoie, What is learned from experience in a probabilistic environment?, J. Behav. Decis. Mak. 17 (2004) 213–229.
[35] D. Ellsberg, Risk, ambiguity, and the Savage axioms, Q. J. Econ. 75 (1961) 643–669.
[36] J.S.B. Evans, Bias in Human Reasoning: Causes and Consequences, Lawrence Erlbaum Associates, Inc., 1989.
[37] J.S.B. Evans, Hypothetical Thinking: Dual Processes in Reasoning and Judgement, vol. 3, Psychology Press, 2007.
[38] J.S.B. Evans, K.E. Stanovich, Dual-process theories of higher cognition: advancing the debate, Perspect. Psychol. Sci. 8 (2013) 223–241.
[39] E. Fantino, J. Kulik, S. Stolarz-Fantino, W. Wright, The conjunction fallacy: a test of averaging hypotheses, Psychon. Bull. Rev. 4 (1997) 96–101.
[40] P.M. Fernbach, A. Darlow, S.A. Sloman, When good evidence goes bad: the weak evidence effect in judgment and decision-making, Cognition 119 (2011) 459–467.
[41] B. Fischoff, Debiasing, Technical Report. Decision Research, Eugene, OR, 1981.
[42] J.E. Fisk, Judgments under uncertainty: representativeness or potential surprise?, Br. J. Psychol. 93 (2002) 431–449.
[43] S.T. Fiske, Attention and weight in person perception: the impact of negative and extreme behavior, J. Pers. Soc. Psychol. 38 (1980) 889.
[44] G.J. Fitzsimons, B. Shiv, Nonconscious and contaminative effects of hypothetical questions on subsequent decision making, J. Consum. Res. 28 (2001) 224–238.
[45] M. Fleischmann, M. Amirpur, A. Benlian, T. Hess, Cognitive biases in information systems research: a scientometric analysis, in: Proceedings of the 22nd European Conference on Information Systems, ECIS 2014, Tel Aviv, Israel, 2014.
[46] D. Fleisig, Adding information may increase overconfidence in accuracy of knowledge retrieval, Psychol. Rep. 108 (2011) 379–392.