



# On cognitive preferences and the plausibility of rule-based models

Johannes Fürnkranz<sup>1</sup> · Tomáš Kliegr<sup>2</sup> · Heiko Paulheim<sup>3</sup>

Received: 10 March 2018 / Revised: 21 April 2019 / Accepted: 19 October 2019 /  
Published online: 24 December 2019  
© The Author(s) 2019

## Abstract

It is conventional wisdom in machine learning and data mining that logical models such as rule sets are more interpretable than other models, and that among such rule-based models, simpler models are more interpretable than more complex ones. In this position paper, we question this latter assumption by focusing on one particular aspect of interpretability, namely the plausibility of models. Roughly speaking, we equate the plausibility of a model with the likeliness that a user accepts it as an explanation for a prediction. In particular, we argue that—all other things being equal—longer explanations may be more convincing than shorter ones, and that the predominant bias for shorter models, which is typically necessary for learning powerful discriminative models, may not be suitable when it comes to user acceptance of the learned models. To that end, we first recapitulate evidence for and against this postulate, and then report the results of an evaluation in a crowdsourcing study based on about 3000 judgments. The results do not reveal a strong preference for simple rules, whereas we can observe a weak preference for longer rules in some domains. We then relate these results to well-known cognitive biases such as the conjunction fallacy, the representative heuristic, or the recognition heuristic, and investigate their relation to rule length and plausibility.

**Keywords** Inductive rule learning · Interpretable models · Cognitive bias

---

Editor: Bart Baesens.

---

✉ Johannes Fürnkranz  
juffi@faw.jku.at

Tomáš Kliegr  
tomas.kliegr@vse.cz

Heiko Paulheim  
heiko@informatik.uni-mannheim.de

<sup>1</sup> Department of Computer Science, Johannes Kepler University Linz, Altenberger Straße 66B, 4040 Linz, Austria

<sup>2</sup> Department of Information and Knowledge Engineering, University of Economics, Prague, nám Winstona Churchilla 4, 13067 Prague, Czech Republic

<sup>3</sup> School of Business Informatics and Mathematics, University of Mannheim, 68159 Mannheim, Germany

# 1 Introduction

In their classical definition of the field, Fayyad et al. (1996) have defined knowledge discovery in databases as “*the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data.*” Research has since progressed considerably in all of these dimensions in a mostly data-driven fashion. The validity of models is typically addressed with predictive evaluation techniques such as significance tests, hold-out sets, or cross validation (Japkowicz and Shah 2011), techniques which are now also increasingly used for pattern evaluation (Webb 2007). The novelty of patterns is typically assessed by comparing their local distribution to expected values, in areas such as novelty detection (Markou and Singh 2003a, b), where the goal is to detect unusual behavior in time series, subgroup discovery (Kralj Novak et al. 2009), which aims at discovering groups of data that have unusual class distributions, or exceptional model mining (Duivesteyn et al. 2016), which generalizes this notion to differences with respect to data models instead of data distributions. The search for useful patterns has mostly been addressed via optimization, where the utility of a pattern is defined via a predefined objective function (Hu and Mojsilovic 2007) or via cost functions that steer the discovery process into the direction of low-cost or high-utility solutions (Elkan 2001). To that end, Kleinberg et al. (1998) formulated a data mining framework based on utility and decision theory.

Arguably, the last dimension, understandability or interpretability, has received the least attention in the literature. The reason why interpretability has rarely been explicitly addressed is that it is often equated with the presence of logical or structured models such as decision trees or rule sets, which have been extensively researched since the early days of machine learning. In fact, much of the research on learning such models has been motivated with their interpretability. For example, Fürnkranz et al. (2012) argue that rules “*offer the best trade-off between human and machine understandability*”. Similarly, it has been argued that rule induction offers a good “mental fit” to decision-making problems (van den Eijkel 1999; Weihs and Sondhauss 2003). Their main advantage is the simple logical structure of a rule, which can be directly interpreted by experts not familiar with machine learning or data mining concepts. Moreover, rule-based models are highly modular, in the sense that they may be viewed as a collection of local patterns (Fürnkranz 2005; Knobbe et al. 2008; Fürnkranz and Knobbe 2010), whose individual interpretations are often easier to grasp than the complete predictive theory. For example, Lakkaraju et al. (2016) argued that rule sets (which they call decision sets) are more interpretable than decision lists because they can be decomposed into individual local patterns.

Only recently, with the success of highly precise but largely inscrutable deep learning models, has the topic of interpretability received serious attention, and several workshops in various disciplines have been devoted to the topic of learning interpretable models at conferences like ICML (Kim et al. 2016, 2017, 2018), NIPS (Wilson et al. 2016; Tosi et al. 2017; Müller et al. 2017) or CHI (Gillies et al. 2016). Moreover, several books on the subject have already appeared, or are in preparation (Jair Escalante et al. 2018; Molnar 2019), funding agencies like DARPA have recognized the need for explainable AI,<sup>1</sup> and the General Data Protection Regulation of the EC includes a “right to explanation”, which may have a strong impact on machine learning and data mining solutions (Piatetsky-Shapiro 2018).

The strength of many recent learning algorithms, most notably deep learning (LeCun et al. 2015; Schmidhuber 2015), feature learning (Mikolov et al. 2013), fuzzy systems (Alonso et al. 2015) or topic modeling (Blei 2012), is that latent variables are formed during the learning

<sup>1</sup> <http://www.darpa.mil/program/explainable-artificial-intelligence>.

process. Understanding the meaning of these hidden variables is crucial for transparent and justifiable decisions. Consequently, visualization of such model components has recently received some attention (Chaney and Blei 2012; Zeiler and Fergus 2014; Rothe and Schütze 2016). Alternatively, some research has been devoted to trying to convert such arcane models to more interpretable rule-based or tree-based theories (Andrews et al. 1995; Craven and Shavlik 1997; Schmitz et al. 1999; Zilke et al. 2016) or to develop hybrid models that combine the interpretability of logic with the predictive strength of statistical and probabilistic models (Besold et al. 2017; Tran and d'Avila Garcez 2018; Hu et al. 2016).

Instead of making the entire model interpretable, methods like LIME (Ribeiro et al. 2016) are able to provide local explanations for inscrutable models, allowing to trade off fidelity to the original model with interpretability and complexity of the local model. In fact, Martens and Provost (2014) report on experiments that illustrate that such local, instance-level explanations are preferable to global, document-level models. An interesting aspect of rule-based theories is that they can be considered as hybrids between local and global explanations (Fürnkranz 2005): A rule set may be viewed as a global model, whereas the individual rule that fires for a particular example may be viewed as a local explanation.

Nevertheless, in our view, many of these approaches fall short in that they take the interpretability of rule-based models for granted. Interpretability is often considered to correlate with complexity, with the intuition that simpler models are easier to understand. Principles like Occam's Razor (Blumer et al. 1987) or Minimum Description Length (MDL) (Rissanen 1978) are commonly used heuristics for model selection, and have shown to be successful in overfitting avoidance. As a consequence, most rule learning algorithms have a strong bias towards simple theories. Despite the necessity of bias for simplicity for overfitting avoidance, we argue in this paper that simpler rules are not necessarily more interpretable, at least not when other aspects of interpretability beyond the mere syntactic readability are considered. This implicit equation of comprehensibility and simplicity was already criticized by, e.g., Pazzani (2000), who argued that *"there has been no study that shows that people find smaller models more comprehensible or that the size of a model is the only factor that affects its comprehensibility."* There are also a few systems that explicitly strive for longer rules, and recent evidence has shed some doubt on the assumption that shorter rules are indeed preferred by human experts. We will discuss the relation of rule complexity and interpretability at length in Sect. 2.

Other criteria than accuracy and model complexity have rarely been considered in the learning process. For example, Gabriel et al. (2014) proposed to consider the semantic coherence of its conditions when formulating a rule. Pazzani et al. (2001) show that rules that respect monotonicity constraints are more acceptable to experts than rules that do not. As a consequence, they modify a rule learner to respect such constraints by ignoring attribute values that generally correlate well with other classes than the predicted class. Freitas (2013) reviews these and other approaches, compares several classifier types with respect to their comprehensibility and points out several drawbacks of model size as a single measure of interpretability.

In his pioneering framework for inductive learning, Michalski (1983) stressed its links with cognitive science, noting that *"inductive learning has a strong cognitive science flavor"*, and postulates that *"descriptions generated by inductive inference bear similarity to human knowledge representations"* with reference to Hintzman (1978), an elementary text from psychology on human learning. Michalski (1983) considers adherence to the comprehensibility postulate to be *"crucial"* for inductive rule learning, yet, as discussed above, it is rarely ever explicitly addressed beyond equating it with model simplicity. Miller (2019) makes an important first step by providing a comprehensive review of what is known in the

social sciences about explanations and discusses these findings in the context of explainable artificial intelligence.

In this paper, we primarily intend to highlight this gap in machine learning and data mining research. In particular, we focus on the *plausibility* of rules, which, in our view, is an important aspect that contributes to interpretability (Sect. 2). In addition to the comprehensibility of a model, which we interpret in the sense that the user can understand the learned model well enough to be able to manually apply it to new data, and its justifiability, which specifies whether the model is in line with existing knowledge, we argue that a good model should also be plausible, i.e., be convincing and acceptable to the user. For example, as an extreme case, a default model that always predicts the majority class is very interpretable, but in most cases not very plausible. We will argue that different models may have different degrees of plausibility, even if they have the same discriminative power. Moreover, we believe that the plausibility of a model is—all other things being equal—not related or in some cases even positively correlated with the complexity of a model.

To that end, we also report the results of a crowdsourcing evaluation of learned rules in four domains (Sect. 3). Overall, the performed experiments are based on nearly 3000 judgments collected from 390 distinct participants. The results show that there is indeed no evidence that shorter rules are preferred by humans. On the contrary, we could observe a preference for longer rules in two of the studied domains (Sect. 4). In the following, we then relate this finding to related results in the psychological literature, such as the conjunctive fallacy (Sect. 5) and insensitivity to sample size (Sect. 6). Section 7 is devoted to a discussion of the relevance of conditions in rules, which may not always have the expected influence on one's preference, in accordance with the recently described weak evidence effect. The remaining sections focus on the interplay of cognitive factors and machine readable semantics: Sect. 8 covers the recognition heuristic, Sect. 9 discusses the effect of semantic coherence on interpretability, and Sect. 10 briefly highlights the lack of methods for learning structured rule-based models.

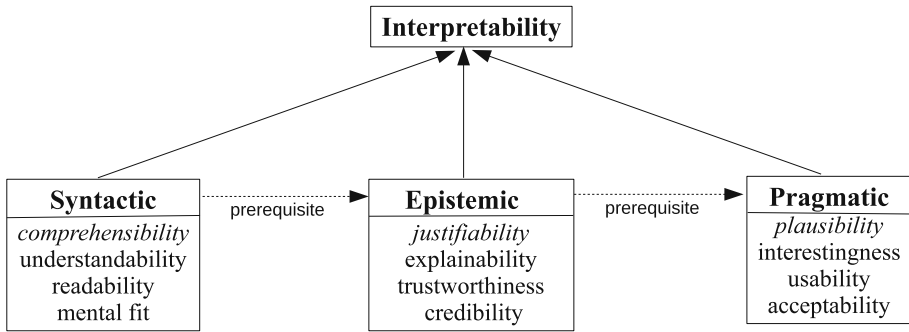
## 2 Aspects of interpretability

Interpretability is a very elusive concept which we use in an intuitive sense. Kodratoff (1994) has already observed that it is an ill-defined concept, and has called upon several communities from both academia and industry to tackle this problem, to “*find objective definitions of what comprehensibility is*”, and to open “*the hunt for probably approximate comprehensible learning*”. Since then, not much has changed. Lipton (2016) still suggests that the term interpretability is ill-defined. In fact, the concept can be found under different names in the literature, including understandability, interpretability, comprehensibility, plausibility, trustworthiness, justifiability and others. They all have slightly different semantic connotations.

A thorough clarification of this terminology is beyond the scope of this paper, but in the following, we briefly highlight different aspects of interpretability and then proceed to clearly define and distinguish comprehensibility and plausibility, the two aspects that are pertinent to this work.

### 2.1 Three aspects of interpretability

In this section, we attempt to bring some order into the multitude of terms that are used in the context of interpretability. Essentially, we distinguish three aspects of interpretability (see also Fig. 1):



**Fig. 1** Three aspects of interpretability

*Syntactic interpretability:* This aspect is concerned with the ability of the user to comprehend the knowledge that is encoded in the model, in very much the same way as the definition of a term can be understood in a conversation or a textbook.

*Epistemic interpretability:* This aspect assesses to what extent the model is in line with existing domain knowledge. A model can be interpretable in the sense that the user can operationalize and apply it, but the encoded knowledge or relationships are not well correlated with the user's prior knowledge. For example, a model which states that the temperature is rising on odd-numbered days and falling on even-numbered days has a high syntactic interpretability but a low epistemic interpretability.

*Pragmatic interpretability:* Finally, we argue that it is important to capture whether the model serves the intended purpose. A model can be perfectly interpretable in the syntactic and epistemic sense but have a low pragmatic value for the user. For example, the simple model that the temperature tomorrow will be roughly the same as today is obviously very interpretable in the syntactic sense, it is also quite consistent with our experience and therefore interpretable in the epistemic sense, but it may not be satisfying as an acceptable explanation for a weather forecast.

Note that these three categories essentially correspond to the grouping of terms pertinent to interpretability, which has previously been introduced by Bibal and Frénay (2016). They treat terms like *comprehensibility*, *understandability*, and *mental fit*, as essentially synonymous to *interpretability*, and use them to denote syntactic interpretability. In a second group, Bibal and Frénay (2016) bring notions such as *interestingness*, *usability*, and *acceptability* together, which essentially corresponds to our notion of pragmatic interpretability. Finally, they have *justifiability* as a separate category, which essentially corresponds to what we mean by epistemic interpretability. We also subsume their notion of *explanatory* as *explainability* in this group, which we view as synonymous to *justifiability*. A key difference to their work lies in our view that all three of the above are different aspects of interpretability, whereas Bibal and Frénay (2016) view the latter two groups as different but related concepts.

We also note in passing that this distinction loosely corresponds to prominent philosophical treatments of explanations (Mayer 2001). Classical theories, such as the deductive-nomological theory of explanation (Hempel and Oppenheim 1948), are based on the validity of the logical connection between premises and conclusion. Instead, Van Fraassen (1977) suggests a pragmatic theory of explanations, according to which the explanation should provide the answer to a (why-)question. Therefore, the same proposition may have different explanations, depending on the information demand. For example, an explanation for why a patient was infected with a certain disease may relate to her medical conditions (for the

doctor) or to her habits (for the patient). Thus, pragmatic interpretability is a much more subjective and user-centered notion than epistemic interpretability.

However, clearly, these aspects are not independent. As already noted by Bibal and Frénay (2016), syntactic interpretability is a prerequisite to the other two notions. We also view epistemic interpretability as a prerequisite to pragmatic interpretability: In case a model is not in line with the user's prior knowledge and therefore has a low epistemic value, it also will have a low pragmatic value to the user. Moreover, the differences between the terms shown in Fig. 1 are soft, and not all previous studies have used them in consistent ways. For example, Muggleton et al. (2018) employ a primarily syntactic notion of comprehensibility (as we will see in Sect. 2.2) and evaluate it by testing whether the participants in their study can successfully apply the acquired knowledge to new problems. In addition, it is also measured whether they can give meaningful names to the explanations they deal with, and whether these names are helpful in applying the knowledge. Thus, these experiments try to capture epistemic aspects as well.

## 2.2 Comprehensibility

One of the few attempts for an operational definition of interpretability is given in the works of Schmid et al. (2017) and Muggleton et al. (2018), who related the concept to objective measurements such as the time needed for inspecting a learned concept, for applying it in practice, or for giving it a meaningful and correct name. This gives interpretability a clearly syntactic interpretation in the sense defined in Sect. 2.1. Following Muggleton et al. (2018), we refer to this type of syntactic interpretability as *comprehensibility* and define it as follows:

**Definition 1** (*Comprehensibility*) A model  $\mathbf{m}_1$  is more “comprehensible” than a model  $\mathbf{m}_2$  with respect to a given task if a human user makes fewer mistakes in the application of model  $\mathbf{m}_1$  to new samples drawn randomly from the task domain than when applying  $\mathbf{m}_2$ .

Thus, a model is considered to be comprehensible if a user is able to understand all the mental calculations that are prescribed by the model, and can successfully apply the model to new tasks drawn from the same population. A model is more comprehensible than another model if the user's error rate in doing so is smaller.<sup>2</sup> Muggleton et al. (2018) study various related, measurable quantities, such as the inspection time, the rate with which the meaning of the predicate is recognized from its definition, or the time used for coming up with a suitable name for a definition.

*Relation to alternative notions of interpretability.* Piltaver et al. (2016) use a very similar definition when they study how the response time for various data- and model-related tasks such as “classify”, “explain”, “validate”, or “discover” varies with changes in the structure of learned decision trees. Another variant of this definition was suggested by Dhurandhar et al. (2017; 2018), who consider interpretability relative to a target model, typically (but not necessarily) a human user. More precisely, they define a learned model as  $\delta$ -interpretable relative to a target model if the target model can be improved by a factor of  $\delta$  (e.g., w.r.t. predictive accuracy) with information obtained by the learned model. All these notions have

<sup>2</sup> We are grateful to one of our reviewers for pointing out that this essentially is in line with the cognitive science perspective on comprehension as proposed by Johnson-Laird (1981), where understanding a natural language sentence or text means to be able to draw valid conclusions and inferences from it.

in common that they relate interpretability to a performance aspect, in the sense that a task can be performed better or performed at all with the help of the learned model.

As illustrated in Fig. 1, we consider understandability, readability and mental fit as alternative terms for syntactic interpretability. Understandability is considered as a direct synonym for comprehensibility (Bibal and Fréney 2016). Readability clearly corresponds to the syntactic level. The term *mental fit* may require additional explanation. We used it in the sense of van den Eijkel (1999) to denote the suitability of the representation (i.e. rules) for a given purpose (to explain a classification model).

## 2.3 Justifiability

A key aspect on interpretability is that a concept is consistent with available domain knowledge, which we call epistemic interpretability. Martens and Baesens (2010) have introduced this concept under the name of *justifiability*. They consider a model to be more justifiable if it better conforms to domain knowledge, which may be viewed as constraints to which a justifiable model has to conform (hard constraints) or should better conform (soft constraints). Martens et al. (2011) provide a taxonomy of such constraints, which include univariate constraints such as monotonicity as well as multivariate constraints such as preferences for groups of variables.

We paraphrase and slightly generalize this notion in the following definition:

**Definition 2** (*Justifiability*) A model  $\mathbf{m}_1$  is more “justifiable” than a model  $\mathbf{m}_2$  if  $\mathbf{m}_1$  violates fewer constraints that are imposed by the user’s prior knowledge.

Martens et al. (2011) also define an objective measure for justifiability, which essentially corresponds to a weighted sum over the fractions of cases where each variable is needed in order to discriminate between different class values.

*Relation to comprehensibility and plausibility.* Definition 1 (comprehensibility) addresses the syntactical level of understanding, which is a prerequisite for justifiability. What this definition does not cover are facets of interpretability that relate to one’s background knowledge. For example, an empty model or a default model, classifying all examples as positive, is very simple to interpret, comprehend and apply, but such model will hardly be justifiable.

Clearly, one needs to be able to comprehend the definition of a concept before it can be checked whether it corresponds to existing knowledge. Conversely, we view justifiability as a prerequisite to our notion of plausibility, which we will define more precisely in the next section: a theory that does not conform to domain knowledge is not plausible, but, on the other hand, the user may nevertheless assess different degrees of plausibility to different explanations that are all consistent with our knowledge. In fact, many scientific and in particular philosophical debates are about different, conflicting theories, which are all justifiable but have different degrees of plausibility for different groups of people.

*Relation to alternative notions of interpretability.* Referring to Fig. 1, we view plausibility as an aspect of epistemic interpretability, similar to notions like explainability, trustworthiness and credibility. Both trustworthiness and credibility imply an evaluation of the model against domain knowledge. Explainability is harder to define and has received multiple definitions in the literature. We essentially follow Gall (2019), who makes a distinction that is similar to our notions of syntactic and epistemic interpretability: in his view, interpretability is to allow the user to grasp the mechanics of a process (similar to the notion of mental fit that we have used above), whereas explainability also implies a deeper understanding of why the process

QOL = High :- Many events take place.  
 QOL = High :- Host City of Olympic Summer Games.  
 QOL = Low :- African Capital.

(a) rated highly by users

QOL = High :- # Records Made  $\geq 1$ , # Companies/Organisations  $\geq 22$ .  
 QOL = High :- # Bands  $\geq 18$ , # Airlines founded in 2000  $> 1$ .  
 QOL = Low :- # Records Made = 0, Average January Temp  $\leq 16$ .

(b) rated lowly by users

**Fig. 2** Good discriminative rules for the quality of living of a city (Paulheim 2012b)

works in this way. This requires the ability to relate the notion to existing knowledge, which is why we view it primarily as an aspect of epistemic interpretability.

## 2.4 Plausibility

In this paper, we focus on a pragmatic aspect of interpretability, which we refer to as *plausibility*. We primarily view this notion in the sense of “user acceptance” or “user preference”. However, as discussed in Sect. 2.1, this also means that it has to rely on aspects of syntactic and epistemic interpretability as prerequisites. For the purposes of this paper, we define plausibility as follows:

**Definition 3** (*Plausibility*) A model  $m_1$  is more “plausible” than a model  $m_2$  if  $m_1$  is more likely to be accepted by a user than  $m_2$ .

Within this definition, the word “accepted” bears the meaning specified by the Cambridge English Dictionary<sup>3</sup> as “generally agreed to be satisfactory or right”.

Our definition of plausibility is less objective than the above definition of comprehensibility because it always relates to the subject’s perception of the utility of a given explanation, i.e., its pragmatic aspect. Plausibility, in our view, is inherently subjective, i.e., it relates to the question how useful a model is perceived by a user. Thus, it needs to be evaluated in introspective user studies, where the users explicitly indicate how plausible an explanation is, or which of two explanations appears to be more plausible. Two explanations that can equally well be applied in practice (and thus have the same syntactic interpretability) and are both consistent with existing knowledge (and thus have the same epistemic interpretability), may nevertheless be perceived as having different degrees of plausibility.

*Relation to comprehensibility and justifiability.* A model may be consistent with domain knowledge, but nevertheless appear implausible. Consider, e.g., the rules shown in Fig. 2, which have been derived by the Explain-a-LOD system (Paulheim and Fürnkranz 2012). The rules provide several possible explanations for why a city has a high quality of living, using Linked Open Data as background knowledge. Clearly, all rules are comprehensible and can be easily applied in practice. They also appear to be justifiable in the sense that all of them appear to be consistent with prior knowledge. For example, while the number of records made in a city is certainly not a *prima facie* aspect of its quality of living, it is reasonable to assume a correlation between these two variables. Nevertheless, the first three rules appear

<sup>3</sup> <https://dictionary.cambridge.org/dictionary/english/accepted>.



to be more plausible to a human user, which was also confirmed in an experimental study (Paulheim 2012a, b).

*Relation to alternative notions of interpretability.* In Fig. 1 we consider *interestingness*, *usability*, and *acceptability* as related terms. All these notions imply some degree of user acceptance or fitness for given purpose.

In the remainder of the paper, we will typically talk about “plausibility” in the sense defined above, but we will sometimes use terms like “interpretability” as a somewhat more general term. We also use “comprehensibility”, mostly when we refer to syntactic interpretability, as discussed and defined above. However, all terms are meant to be interpreted in an intuitive, and non-formal way.<sup>4</sup>

### 3 Setup of crowdsourcing experiments on plausibility

In the remainder of the paper, we focus on the plausibility of rules. In particular, we report on a series of five crowdsourcing experiments, which relate the perceived plausibility of a rule to various factors such as rule complexity, attribute importance or centrality. As a basis, we used pairs of rules generated by machine learning systems, typically one rule representing a shorter, and the other a longer explanation. Participants were then asked to indicate which one of the pair they preferred.

The selection of crowdsourcing as a means of acquiring data allows us to gather thousands of responses in a manageable time frame while at the same time ensuring our results can be easily replicated.<sup>5</sup> In the following, we describe the basic setup that is common to all performed experiments. Most of the setup is shared for the subsequent experiments and will not be repeated, only specific deviations will be mentioned. Cognitive science research has different norms for describing experiments than those that are commonly employed in machine learning research.<sup>6</sup> Also, the parameters of the experiments, such as the amount of payment, is described in somewhat greater detail than usual in machine learning, because of the general sensitivity of the participants to such conditions.

We tried to respect these differences by dividing experiment descriptions here and in subsequent sections into subsections entitled “*Material*”, “*Participants*”, “*Methodology*”, and “*Results*”, which correspond to the standard outline of an experimental account in cognitive science. In the following, we describe the general setup that applies to all experiments in the following sections, where then the main focus can be put on the results.

---

<sup>4</sup> In particular, we do not intend to touch upon formal notions of plausibility, such as those given in the Dempster-Shafer theory, where plausibility of an evidence is defined as an upper bound on the belief in the evidence, or more precisely, as the converse of one’s belief in the opposite of the evidence (Dempster 1967; Shafer 1976).

<sup>5</sup> To this end, source datasets, preprocessing code, the responses obtained with crowdsourcing, and the code used to analyze them were made available at <https://github.com/kliegr/rule-length-project>. The published data do not contain quiz failure rates (*qfr* in Tables 4–8) since these were computed from statistics only displayed in the dashboard of the used crowdsourcing platform upon completion of the crowdsourcing tasks.

<sup>6</sup> In fact, with *psychometrics*, an entire field is devoted to proper measurement of psychological phenomena (Furr and Bacharach 2008).

**Table 1** Overview of the datasets used for generating rule pairs

# Pairs	Dataset	Data source	# Instances	# attr.	Target
80	<i>Traffic</i>	LOD	146	210	Rate of traffic accidents in a country
36	<i>Quality</i>	LOD	230	679	Quality of living in a city
32	<i>Movies</i>	LOD	2000	1770	Movie rating
10	<i>Mushroom</i>	UCI	8124	23	Mushroom poisonous/edible

### 3.1 Material

For each experiment, we generated rule pairs generated with two different learning algorithms, and asked users about their preference. The details of the rule generation and selection process are described in this section.

#### 3.1.1 Domains

For the experiment, we used learned rules in four domains (Table 1):

- Mushroom* contains mushroom records drawn from Field Guide to North American Mushrooms (Lincoff 1981). Being available at the UCI repository (Dua and Karra Taniskidou 2017), it is arguably one of the most frequently used datasets in rule learning research, its main advantage being discrete, understandable attributes.
- Traffic* is a statistical dataset of death rates in traffic accidents by country, obtained from the WHO.<sup>7</sup>
- Quality* is a dataset derived from the Mercer Quality of Living index, which collects the perceived quality of living in cities world wide.<sup>8</sup>
- Movies* is a dataset of movie ratings obtained from MetaCritic.<sup>9</sup>

The last three datasets were derived from the Linked Open Data (LOD) cloud (Ristoski et al. 2016). Originally, they consisted only of a name and a target variable, such as a city and its quality-of-living index, or a movie and its rating. The names were then linked to entities in the public LOD dataset DBpedia, using the method described by Paulheim and Fürnkranz (2012). From that dataset, we extracted the classes to which the entities belong, using the deep classification of YAGO, which defines a very fine grained class hierarchy of several thousand classes. Each class was added as a binary attribute. For example, the entity for the city of Vienna would get the binary features *European Capitals*, *UNESCO World Heritage Sites*, etc.

The goal behind these selections was that the domains are general enough so that the participants are able to comprehend a given rule without the need for additional background knowledge, but are nevertheless not able to reliably judge the validity of a given rule. Thus, participants will need to rely on their common sense in order to judge which of the two rules appears to be more convincing. This also implies that we specifically did not expect the users to have expert knowledge in these domains.

<sup>7</sup> [http://www.who.int/violence\\_injury\\_prevention/road\\_traffic/en/](http://www.who.int/violence_injury_prevention/road_traffic/en/).

<sup>8</sup> <http://across.co.nz/qualityofliving.htm>.

<sup>9</sup> <http://www.metacritic.com/movie>.

### 3.1.2 Rule generation

We used two different approaches to generate rules for each of the four domains mentioned in the previous section.

*Class Association Rules:* We used a standard implementation of the APRIORI algorithm for association rule learning (Agrawal et al. 1993; Hahsler et al. 2011) and filtered the output for class association rules with a minimum support of 0.01, minimum confidence of 0.5, and a maximum length of 5. Pairs were formed between all rules that correctly classified at least one shared instance. Although other more sophisticated approaches (such as a threshold on the Dice coefficient) were considered, it turned out that the process outlined above produced rule pairs with quite similar values of confidence (i.e., most equal to 1.0), except for the *Movies* dataset.

*Classification Rules:* We used a simple top-down greedy hill-climbing algorithm that takes a seed example and generates a pair of rules, one with a regular heuristic (Laplace) and one with its inverted counterpart. As shown by Stecher et al. (2016) and illustrated in Fig. 5, this results in rule pairs that have approximately the same degree of generality but different complexities.

From the resulting rule sets, we selected several *rule pairs* consisting of a long and a short rule that have the same or a similar degree of generality.<sup>10</sup> For *Quality* and *Movies*, all rule pairs were used. For the *Mushroom* dataset, we selected rule pairs so that every difference in length (one to five) is represented. All selected rule pairs were pooled, so we did not discriminate between the learning algorithm that was used for generating them. For the *Traffic* dataset, the rule learner generated a higher number of rules than for the other datasets, which allowed us to select the rule pairs for annotation in such a way that various types of differences between rules in each pair were represented. Since this stratification procedure, detailed in (Kliegr 2017), applied only to one of the datasets, we do not expect this design choice to have a profound impact on the overall results and omit a detailed description here.

As a final step, we automatically translated all rule pairs into human-friendly HTML-formatted text, and randomized the order of the rules in the rule pair. Example rules for the four datasets are shown in Fig. 3. The first column of Table 1 shows the final number of rule pairs generated in each domain.

## 3.2 Methodology

The generated rule pairs were then evaluated in a user study on a crowdsourcing platform, where participants were asked to issue a preference between the plausibility of the shown rules. This was then correlated to various factors that could have an influence on plausibility.

---

<sup>10</sup> The generality of a rule is defined via the set of examples a rule covers. Two rules that cover the same *set* of examples have the same generality, even if they have a different number of conditions. Examples include an itemset and its closure, or the elephant example discussed further below in Sect. 4.2. We use the phrase “degree of generality” somewhat loosely to refer to two rules that cover an equal *number* of examples, such as the pair of first rules of the two solutions in the *Mushroom* dataset (Fig. 5).

<p>if the country falls into all of the following groups simultaneously</p> <ul style="list-style-type: none"> <li>* European Union Member Economies and</li> <li>* Imperial free cities</li> </ul> <p>then the quality of living is <b>highest</b></p>	<p>if the movie falls into all of the following group(s) (simultaneously)</p> <ul style="list-style-type: none"> <li>* Films Released in 2005 and</li> <li>* Englishlanguage Films</li> </ul> <p>then the movie is rated as <b>good</b></p>
<p>if the country falls into all of the following groups simultaneously</p> <ul style="list-style-type: none"> <li>* States And Territories Established In 2006 and</li> <li>* Serbo-Croatian-Speaking Countries</li> </ul> <p>then the risk of traffic accidents is <b>low</b></p>	<p>if the mushroom falls into all of the following groups simultaneously</p> <ul style="list-style-type: none"> <li><i>veil color is white and</i></li> <li><i>stalk surface below ring is silky</i></li> </ul> <p>then the mushroom is <b>poisonous</b></p>

**Fig. 3** Example translated rules for the four datasets

### 3.2.1 Definition of crowdsourcing experiments

As the experimental platform we used the CrowdFlower crowdsourcing service.<sup>11</sup> Similar to the better-known Amazon Mechanical Turk, CrowdFlower allows to distribute questionnaires to participants around the world, who complete them for remuneration. The remuneration is typically a small payment in US dollars—for one judgment relating to one rule we paid 0.07 USD—but some participants may receive the payment in other currencies, including in game currencies (“coins”).

A crowdsourcing task performed in CrowdFlower consists of a sequence of steps:

1. The CrowdFlower platform recruits participants, so-called *workers* for the task from a pool of its users, who match the level and geographic requirements set by the experimenter. The workers decide to participate in the task based on the payment offered and the description of the task.
2. Participants are presented assignments which contain an illustrative example.
3. If the task contains test questions, each worker has to pass a *quiz mode* with test questions. Participants learn about the correct answer after they pass the quiz mode, and have the option to contest the correct answer if they consider it incorrect.
4. Participants proceed to the *work mode*, where they complete the task they have been assigned by the experimenter. The task typically has the form of a questionnaire. If test questions were defined by the experimenter, the CrowdFlower platform randomly inserts test questions into the questionnaire. Failing a predefined proportion of hidden test questions results in removal of the worker from the task. Failing the initial quiz or failing a task can also reduce participants’ accuracy on the CrowdFlower platform. Based on the average accuracy, participants can reach one of the three levels. A higher level gives a user access to additional, possibly better paying tasks.
5. Participants can leave the experiment at any time. To obtain payment for their work, they need to submit at least one page of work. After completing each page of work, the worker can opt to start another page. The maximum number of pages per participant is set by the experimenter. As a consequence, two workers can contribute a different number of judgments to the same task.
6. If a bonus was promised, the qualifying participants receive extra credit.

<sup>11</sup> Since our experiments, CrowdFlower has been re-branded under the name *Figure Eight* and is now available at <https://www.figure-eight.com/>.

**Overview**

We kindly ask you to assist us in an experiment which will help to understand how humans perceive rules describing data.

Example rule is

If the mushroom falls into all of the following groups simultaneously

- \* veil color is *white* and
- \* stalk surface below ring is *silky*

then the mushroom is **poisonous**

**Your task**

Your task is to assess which of the two rules in each pair is more **plausible**.

You are not required to factually check the correctness of the rules. What we are interested in is your perception of their **plausibility**.

WARNING: Note that the presented rules may not necessarily be correct.

**PLEASE MAKE SURE TO READ AND UNDERSTAND THE EXPLANATION OF PLAUSIBILITY BELOW BEFORE YOU START THE TASK**

I assess rule plausibility, what is it?

When assessing plausibility, please follow these dictionary definitions:

- (Of an argument or statement) seeming reasonable or probable.
- Seeming likely to be true, or able to be believed
- Possibly true; able to be believed.

Obtaining further information

If the meaning of one of the conditions in the rule is not clear, you can search for explanation on [Wikipedia](#).

**Thank you for your assistance !**

**Fig. 4** Example instructions for experiments 1–3. The example rule pair was adjusted based on the dataset. For Experiment 3, the box with the example rule additionally contained values of confidence and support, formatted as shown in Fig. 8

The workers were briefed with task instructions, which described the purpose of the task, gave an example rule, and explained plausibility as the elicited quantity (cf. Fig. 4). As part of the explanation, the participants were given definitions of “plausible” sourced from the Oxford Dictionary<sup>12</sup> and the Cambridge Dictionary<sup>13</sup> (British and American English). The individual task descriptions differed for the five tasks and will be described in more detail later in the paper in the corresponding sections. Table 2 shows a brief overview of the factors variables and their data types for the five experiments.

<sup>12</sup> <https://en.oxforddictionaries.com/definition/plausible>.

<sup>13</sup> <https://dictionary.cambridge.org/dictionary/english/plausible>.

**Table 2** Variables used in Experiment 1–5

Exp.	Independent variable		Dependent variable	
	Name	Data type	Name	Data type
1	Rule length $\Delta$	Continuous	Plausibility	Discrete
2	Rule length $\Delta$	Continuous	Plausibility	Discrete
3	Rule support $\Delta$ , rule confidence $\Delta$	Discrete	Plausibility	Discrete
4	Attribute importance avg $\Delta$ , att. imp. max $\Delta$ , literal imp. avg $\Delta$ , lit. imp. max $\Delta$	Continuous	Plausibility	Discrete
5	Literal PageRank avg $\Delta$ , literal PageRank max $\Delta$	Continuous	Plausibility	Discrete

$\Delta$  after a variable refers to the difference of its values for a given rule pair. Plausibility was elicited on a five-level linguistic scale ranging from  $-2$  for “Rule 2 (strong preference)” to  $+2$  for “Rule 1 (strong preference)”

### 3.2.2 Evaluation

Rule plausibility was elicited on a five-level linguistic scale ranging from “Rule 2 (strong preference)” to “Rule 1 (strong preference)”, which were interpreted as ordinal values from  $-2$  to  $+2$ . Evaluations were performed at the level of individual judgments, also called micro-level, i.e., each response was considered to be a single data point, and multiple judgments for the same pair were not aggregated prior to the analysis. By performing the analysis at the micro-level, we avoided the possible loss of information as well as the aggregation bias (Clark and Avery 1976). Also, as shown, for example, by Robinson (1950), the ecological (macro-level) correlations are generally larger than the micro-level correlations, therefore by performing the analysis on the individual level, we obtain more conservative results.

We report rank correlation between a factor and the observed evaluation (Kendall’s  $\tau$ , Spearman’s  $\rho$ ) and tested whether the coefficients are significantly different from zero. We will refer to the values of Kendall’s  $\tau$  as the primary measure of rank correlation, since according to Kendall and Gibbons (1990) and Newson (2002), the confidence intervals for Spearman’s  $\rho$  are less reliable than confidence intervals for Kendall’s  $\tau$ .

For all obtained correlation coefficients we compute the  $p$ -value, which is the probability of obtaining a correlation coefficient at least as extreme as the one that was actually observed assuming that the null hypothesis holds, i.e., that there is no correlation between the two variables. The typical cutoff value for rejecting the null hypothesis is  $\alpha = 0.05$ .

### 3.3 Participants

The workers in the CrowdFlower platform were invited to participate in individual tasks. CrowdFlower divides the available workforce into three levels depending on the accuracy they obtained on earlier tasks. As the level of the CrowdFlower workers, we chose Level 2, which was described as follows: “Contributors in Level 2 have completed over a hundred Test Questions across a large set of Job types, and have an extremely high overall Accuracy.”

In order to avoid spurious answers, we also employed a minimum threshold of 180 s for completing a page; workers taking less than this amount of time to complete a page were removed from the job. The maximum time required to complete the assignment was not specified, and the maximum number of judgments per contributor was not limited.

**Table 3** Geographical distribution of collected judgments

	Group 1 judgments			Group 2 judgments			Group 3 judgments			Total judgments
	USA	UK	Can	USA	UK	Can	USA	UK	Can	
(a) Experiments 1–3										
<i>Quality</i>	68	64	52	96	40	44				364
<i>Movies</i>	80	52	28	76	30	58	84	44	32	484
<i>Traffic</i>	204	120	84	212	116	72				808
<i>Mushroom</i>	106	84	60	97	21	32				400
<b>Total</b>	<b>458</b>	<b>320</b>	<b>224</b>	<b>481</b>	<b>207</b>	<b>206</b>	<b>84</b>	<b>44</b>	<b>32</b>	<b>2056</b>
	Literal relevance			Attribute relevance						Total judgments
	USA	UK	Can	USA	UK	Can				
(b) Experiments 4–5										
<i>Quality</i>	63	65	37							165
<i>Movies</i>	74	46	30							150
<i>Traffic</i>	164	58	68	0	10	25				325
<i>Mushroom</i>	70	56	44	23	31	38				262
<b>Total</b>	<b>371</b>	<b>225</b>	<b>179</b>	<b>23</b>	<b>41</b>	<b>63</b>				<b>902</b>

For quality assurance, each participant who decided to accept the task first faced a quiz consisting of a random selection of previously defined test questions. These had the same structure as regular questions but additionally contained the expected correct answer (or answers) as well as an explanation for the answer. We used *swap test* questions where the order of the conditions was randomly permuted in each of the two pairs so that the participant should not have a preference for either of the two versions. The correct answer and explanation was only shown after the worker had responded to the question. Only workers achieving at least 70% accuracy on test questions could proceed to the main task.

### 3.3.1 Statistical information about participants

CrowdFlower does not publish demographic data about its base of workers. Nevertheless, for all executed tasks, the platform makes available the location of the worker submitting each judgment. In this section, we use this data to elaborate on the number and geographical distribution of workers participating in Experiments 1–5 described later in this paper.

Table 3a reports on workers participating in Experiments 1–3, where three types of guidelines were used in conjunction with four different datasets, resulting in 9 tasks in total (not all combinations were tried). Experiments 4–5 involved different guidelines (for determining attribute and literal relevance) and the same datasets. The geographical distribution is reported in Table 3b. In total, the reported results are based on 2958 *trusted* judgments.<sup>14</sup> Actually, more judgments were collected, but some were excluded due to automated quality checks.

<sup>14</sup> A trusted judgment is an answer from a worker that passed the initial quiz and on submitting the work had accuracy score higher than minimum preset accuracy on any hidden test questions. Only trusted judgments were used for analyses.

In order to reduce possible effects of language proficiency, we restricted our participants to English-speaking countries. Most judgments (1417) were made by workers from United States, followed by the United Kingdom (837) and Canada (704). The number of distinct participants for each crowdsourcing task is reported in detailed tables describing the results of the corresponding experiments (*part* column in Tables 4–8). Note that some workers participated in multiple tasks. The total number of distinct participants across all tasks reported in Tables 3a and 3b is 390.

### 3.3.2 Representativeness of crowdsourcing experiments

There is a number of differences between crowdsourcing and the controlled laboratory environment previously used to run psychological experiments. The central question is to what extent do the cognitive abilities and motivation of participants differ between the crowdsourcing cohort and the controlled laboratory environment. Since there is a small amount of research specifically focusing on the population of the CrowdFlower platform, which we use in our research, we present data related to Amazon Mechanical Turk, under the assumption that the descriptions of the populations will not differ substantially.<sup>15</sup> This is also supported by previous work such as Wang et al. (2015), which has indicated that the user distribution of CrowdFlower and AMT is comparable.

The population of crowdsourcing workers is a subset of the population of Internet users, which is described in a recent meta study by Paolacci and Chandler (2014) as follows: “*Workers tend to be younger (about 30 years old), overeducated, underemployed, less religious, and more liberal than the general population.*” While there is limited research on workers’ cognitive abilities, Paolacci et al. (2010) found “*no difference between workers, undergraduates, and other Internet users on a self-report measure of numeracy that correlates highly with actual quantitative abilities.*” According to a more recent study by Crump et al. (2013), workers learn more slowly than university students and may have difficulties with complex tasks. Possibly the most important observation related to the focus of our study is that according to Paolacci et al. (2010) crowdsourcing workers “*exhibit the classic heuristics and biases and pay attention to directions at least as much as subjects from traditional sources.*”

## 4 Interpretability, plausibility, and model complexity

The rules shown in Fig. 2 may suggest that simpler rules are more acceptable than longer rules because the highly rated rules (a) are shorter than the lowly rated rules (b). In fact, there are many good reasons why simpler models should be preferred over more complex models. Obviously, a shorter model can be interpreted with less effort than a more complex model of the same kind, in much the same way as reading one paragraph is quicker than reading one page. Nevertheless, a page of elaborate explanations may be more comprehensible than a single dense paragraph that provides the same information (as we all know from reading research papers).

Other reasons for preferring simpler models include that they are easier to falsify, that there are fewer simpler theories than complex theories, so the a priori chances that a simple theory

---

<sup>15</sup> This is supported by the fact that until about 2014, CrowdFlower platform involved Amazon Mechanical Turk (AMT) workers. As of 2017, these workers are no longer involved, because according to CrowdFlower, the AMT channel was both slower and less accurate than other channels used by the CrowdFlower platform (cf. <http://turkrequesters.blogspot.com/2014/01/crowdfower-dropping-mechanical-turk.html>).



fits the data are lower, or that simpler rules tend to be more general, cover more examples and their quality estimates are therefore statistically more reliable.

However, one can also find results that throw doubt on this claim. In particular, in cases where not only syntactic interpretability is considered, there are some previous works where it was observed that longer rules are preferred by human experts. In the following, we discuss this issue in some depth, by first reviewing the use of a simplicity bias in machine learning (Sect. 4.1), then taking the alternative point of view and recapitulating works where more complex theories are preferred (Sect. 4.2), and then summarizing the conflicting past evidence for either of the two views (Sect. 4.3). Finally, in Sect. 4.4, we report on the results of our first experiment, which aimed at testing whether rule length has an influence on the interpretability or plausibility of found rules at all, and, if so, whether people tend to prefer longer or shorter rules.

#### 4.1 The bias for simplicity

Michalski (1983) already states that inductive learning algorithms need to incorporate a preference criterion for selecting hypotheses to address the problem of the possibly unlimited number of hypotheses and that this criterion is typically *simplicity*, referring to philosophical works on simplicity of scientific theories by Kemeny (1953) and Post (1960), which refine the initial postulate attributed to Ockham, which we discuss further below. According to Post (1960), judgments of simplicity should not be made “*solely on the linguistic form of the theory*”.<sup>16</sup> This type of simplicity is referred to as *linguistic simplicity*. A related notion of *semantic simplicity* is described through the falsifiability criterion (Popper 1935, 1959), which essentially states that simpler theories can be more easily falsified. Third, Post (1960) introduces *pragmatic simplicity*, which relates to the degree to which the hypothesis can be fitted into a wider context.

Machine learning algorithms typically focus on linguistic or syntactic simplicity, by referring to the description length of the learned hypotheses. The complexity of a rule-based model is typically measured with simple statistics, such as the number of learned rules and their length, or the total number of conditions in the learned model (cf., e.g., Todorovski et al. 2000; Lakkaraju et al. 2016; Minnaert et al. 2015; Wang et al. 2017). Inductive rule learning is typically concerned with learning a set of rules or a rule list that discriminates positive from negative examples (Fürnkranz et al. 2012; Fürnkranz and Klieger 2015). For this task, a bias towards simplicity is necessary because, for a contradiction-free training set, it is trivial to find a rule set that perfectly explains the training data, simply by converting each example to a maximally specific rule that covers only this example.

Occam’s Razor, “*Entia non sunt multiplicanda sine necessitate*”,<sup>17</sup> which is attributed to English philosopher and theologian William of Ockham (c. 1287–1347), has been put forward as support for a principle of parsimony in the philosophy of science (Hahn 1930). In machine learning, this principle is generally interpreted as “*given two explanations of the data, all other things being equal, the simpler explanation is preferable*” (Blumer et al. 1987), or simply “*choose the shortest explanation for the observed data*” (Mitchell 1997). While it is well-known that striving for simplicity often yields better predictive results—mostly because pruning or regularization techniques help to avoid overfitting—the exact

<sup>16</sup> Kemeny (1953) gave the example that among competing explanations for the solar system, the model of Tycho Brahe is linguistically simpler than Copernicus’ theory because of the convenient choice of the coordinate system associated with the heliocentric view.

<sup>17</sup> Entities should not be multiplied beyond necessity.

formulation of the principle is still subject to debate (Domingos 1999), and several cases have been observed where more complex theories perform better (Murphy and Pazzani 1994; Webb 1996; Bensusan 1998).

Much of this debate focuses on the aspect of predictive accuracy. When it comes to understandability, the idea that simpler rules are more comprehensible is typically unchallenged. A nice counterexample is due to Munroe (2013), who observed that route directions like “*take every left that doesn’t put you on a prime-numbered highway or street named for a president*” could be most compressive but considerably less comprehensive. Although Domingos (1999) argues in his critical review that it is theoretically and empirically false to favor the simpler of two models with the same training-set error on the grounds that this would lead to lower generalization error, he concludes that Occam’s Razor is nevertheless relevant for machine learning but should be interpreted as a preference for more *comprehensible* (rather than *simple*) models. Here, the term “comprehensible” clearly does not refer to syntactical length.

A particular implementation of Occam’s razor in machine learning is the minimum description length (MDL; Rissanen 1978) or minimum message length (MML<sup>18</sup>; Wallace and Boulton 1968) principle which is an information-theoretic formulation of the principle that smaller models should be preferred (Grünwald 2007). The description length that should be minimized is the sum of the complexity of the model plus the complexity of the data encoded given the model. In this way, both the complexity and the accuracy of a model can be traded off: the description length of an empty model consists only of the data part, and it can be compared to the description length of a perfect model, which does not need additional information to encode the data. The theoretical foundation of this principle is based on the Kolmogorov complexity (Li and Vitányi 1993), the essentially uncomputable length of the smallest model of the data. In practice, different coding schemes have been developed for encoding models and data and have, e.g., been used as selection or pruning criteria in decision tree induction (Needham and Dowe 2001; Mehta et al. 1995), inductive rule learning (Quinlan 1990; Cohen 1995; Pfahringer 1995) or for pattern evaluation (Vreeken et al. 2011). The ability to compress information has also been proposed as a basis for human comprehension and thus forms the backbone of many standard intelligence tests, which aim at detecting patterns in data. Psychometric artificial intelligence (Bringsjord 2011) extends this definition to AI in general. For an extensive treatment of the role of compression in measuring human and machine intelligence, we refer the reader to Hernández-Orallo (2017).

Many works make the assumption that the interpretability of a rule-based model can be measured by measures that relate to the complexity of the model, such as the number of rules or the number conditions. A maybe prototypical example is the Interpretable Classification Rule Mining (ICRM) algorithm, which “*is designed to maximize the comprehensibility of the classifier by minimizing the number of rules and the number of conditions*” via an evolutionary process (Cano et al. 2013). Similarly, Minnaert et al. (2015) investigate a rule learner that is able to optimize multiple criteria and evaluate it by investigating the Pareto front between accuracy and comprehensibility, where the latter is measured with the number of rules. Lakkaraju et al. (2016) propose a method for learning rule sets that simultaneously optimizes accuracy and interpretability, where the latter is again measured by several conventional data-driven criteria such as rule overlap, coverage of the rule set, and the number of conditions and rules in the set. Most of these works clearly focus on syntactic interpretability.

<sup>18</sup> The differences between the two views are irrelevant for our argumentation.

## 4.2 The bias for complexity

Even though most systems have a bias toward simpler theories for the sake of overfitting avoidance and increased accuracy, some rule learning algorithms strive for more complex rules and have good reasons for doing so. Already Michalski (1983) has noted that there are two different kinds of rules, discriminative and characteristic. *Discriminative rules* can quickly discriminate an object of one category from objects of other categories. A simple example is the rule

elephant :- trunk.

which states that an animal with a trunk is an elephant. This implication provides a simple but effective rule for recognizing elephants among all animals. However, it does not provide a very clear picture of the properties of the elements of the target class. For example, from the above rule, we do not understand that elephants are also very large and heavy animals with a thick gray skin, tusks and big ears.

*Characteristic rules*, on the other hand, try to capture *all* properties that are common to the objects of the target class. A rule for characterizing elephants could be

heavy, large, gray, bigEars, tusks, trunk :- elephant.

Note that here the implication sign is reversed: we list all properties that are implied by the target class, i.e., by an animal being an elephant. Even though discriminative rules are easier to comprehend in the syntactic sense, we argue that characteristic rules are often more interpretable than discriminative rules from a pragmatic point of view. For example, in a customer profiling application, we might prefer to not only list a few characteristics that discriminate one customer group from the other, but are interested in all characteristics of each customer group.

The distinction between characteristic and discriminative rule is also reminiscent of the distinction between defining and characteristic features of categories. Smith et al. (1974)<sup>19</sup> argue that both of them are used for similarity-based assessments of categories to objects, but that only the defining features are eventually used when similarity-based categorization over all features does not give a conclusive positive or negative answer.

Characteristic rules are very much related to *formal concept analysis* (Wille 1982; Ganter and Wille 1999). Informally, a concept is defined by its intent (the description of the concept, i.e., the conditions of its defining rule) and its extent (the instances that are covered by these conditions). A *formal concept* is then a concept where the extension and the intension are Pareto-maximal, i.e., a concept where no conditions can be added without reducing the number of covered examples. In Michalski's terminology, a formal concept is both discriminative and characteristic, i.e., a rule where the head is equivalent to the body.

It is well-known that formal concepts correspond to *closed itemsets* in association rule mining, i.e., to maximally specific itemsets (Stumme et al. 2002). Closed itemsets have been primarily mined because they are a unique and compact representative of equivalence classes of itemsets, which all cover the same instances (Zaki and Hsiao 2002). However, while all itemsets in such an equivalence class are equivalent with respect to their support, they may not be equivalent with respect to their understandability or interestingness.

Gamberger and Lavrač (2003) introduce *supporting factors* as a means for complementing the explanation delivered by conventional learned rules. Essentially, they are additional

<sup>19</sup> Many thanks to the anonymous reviewer who pointed us towards this and some of the following works.

```

[2160|0] p :- odor = foul.
[1152|0] p :- gill-color = buff.
[ 256|0] p :- odor = pungent.
[ 192|0] p :- odor = cresote.
[  72|0] p :- spore-print-color = green.
[  36|0] p :- stalk-color-below-ring = cinnamon.
[  24|0] p :- stalk-color-below-ring = scaly.
[   4|0] p :- cap-surface = grooves.
[   1|0] p :- cap-shape = conical.
[  16|0] p :- stalk-color-below-ring = brown, stalk-surface-above-ring = silky.
[   3|0] p :- habitat = leaves, stalk-color-below-ring = white.

```

(a) using the Laplace heuristic  $h_{Lap}$  for refinement

```

[2192|0] p :- veil-color = white, gill-spacing = close, bruises? = no,
           ring-number = one, stalk-surface-above-ring = silky.
[ 864|0] p :- veil-color = white, gill-spacing = close, gill-size = narrow,
           population = several, stalk-shape = tapering.
[ 336|0] p :- stalk-color-below-ring = white, ring-type = pendant,
           stalk-color-above-ring = white, ring-number = one,
           cap-surface = smooth, stalk-root = bulbous, gill-spacing = close.
[ 264|0] p :- stalk-surface-below-ring = smooth, stalk-surface-above-ring = smooth,
           ring-type = pendant, stalk-shape = enlarging, veil-color = white,
           gill-size = narrow, bruises? = no.
[ 144|0] p :- stalk-shape = enlarging, stalk-root = bulbous,
           stalk-color-below-ring = white, ring-number = one.
[  72|0] p :- stalk-shape = enlarging, gill-spacing = close, veil-color = white,
           gill-size = broad, spore-print-color = green.
[  44|0] p :- stalk-surface-below-ring = scaly, stalk-root = club.

```

(b) using the inverted Laplace heuristic  $u_{Lap}$  for refinement

**Fig. 5** Two decision lists learned for the class *poisonous* in the *Mushroom* dataset

attributes that are not part of the learned rule, but nevertheless have very different distributions with respect to the classes of the application domain. In a way, enriching a rule with such supporting factors is quite similar to computing the closure of a rule. In line with the results of Kononenko (1993), medical experts found that these supporting factors increase the plausibility of the found rules.

Stecher et al. (2014) introduced so-called *inverted heuristics* for inductive rule learning. The key idea behind them is a rather technical observation based on a visualization of the behavior of rule learning heuristics in coverage space (Fürnkranz and Flach 2005), namely that the evaluation of rule refinements is based on a bottom-up point of view, whereas the refinement process proceeds top-down, in a general-to-specific fashion. As a remedy, it was proposed to “invert” the point of view, resulting in heuristics that pay more attention to maintaining high coverage on the positive examples, whereas conventional heuristics focus more on quickly excluding negative examples. Somewhat unexpectedly, it turned out that this results in longer rules, which resemble characteristic rules instead of the conventionally learned discriminative rules. For example, Fig. 5 shows the two decision lists that have been found for the *Mushroom* dataset with the conventional Laplace heuristic  $h_{Lap}$  (top) and its inverted counterpart  $u_{Lap}$  (bottom). Although fewer rules are learned with  $u_{Lap}$ , and thus the individual rules are more general on average, they are also considerably longer. Intuitively, these rules also look more convincing, because the first set of rules often only uses a single criterion (e.g., odor) to discriminate between edible and poisonous mushrooms. Thus, even though the shorter rules may be more comprehensible in the syntactic sense, the longer rules

appear to be more plausible. Stecher et al. (2016) and Valmarska et al. (2017) investigated the suitability of such rules for subgroup discovery, with somewhat inconclusive results.

### 4.3 Conflicting evidence

The above-mentioned examples should help to motivate that the complexity of models may have an effect on the interpretability of a model. Even in cases where a simpler and a more complex rule covers the same number of examples, shorter rules are not necessarily more interpretable, at least not when other aspects of interpretability beyond syntactic comprehensibility are considered. There are a few isolated empirical studies that add to this picture. However, the results on the relation between the size of representation and interpretability are limited and conflicting, partly because different aspects of interpretability are not clearly discriminated.

*Larger models are less interpretable.* Huysmans et al. (2011) were among the first that actually tried to empirically validate the often implicitly made claim that smaller models are more interpretable. In particular, they related increased complexity to measurable events such as a decrease in answer accuracy, an increase in answer time, and a decrease in confidence. From this, they concluded that smaller models tend to be more interpretable, proposing that there is a certain complexity threshold that limits the practical utility of a model. However, they also noted that in parts of their study, the correlation of model complexity with utility was less pronounced. The study also does not report whether the participants of their study had any domain knowledge relating to the used data, so that it cannot be ruled out that the obtained result was caused by lack of domain knowledge.<sup>20</sup> A similar study was later conducted by Piltaver et al. (2016), who found a clear relationship between model complexity and interpretability in decision trees.

In most previous works, interpretability was interpreted in the sense of syntactic comprehensibility, i.e., the pragmatic or epistemic aspects of interpretability were not addressed.

*Larger models are more interpretable.* A direct evaluation of the perceived interpretability of classification models has been performed by Allahyari and Lavesson (2011). They elicited preferences on pairs of models, which were generated from two UCI datasets: *Labor* and *Contact Lenses*. What is unique to this study is that the analysis took into account the participants' estimated knowledge about the domain of each of the datasets. On *Labor*, they were expected to have good domain knowledge but not so for *Contact Lenses*. The study was performed with 100 students and involved several decision tree induction algorithms (J48, RIDOR, ID3) as well as rule learners (PRISM, REP, JRIP). It was found that *larger models* were considered as *more comprehensible* than smaller models on the *Labor* dataset, whereas the users showed the opposite preference for *Contact Lenses*. Allahyari and Lavesson (2011) explain the discrepancy with the lack of prior knowledge for *Contact Lenses*, which makes it harder to understand complex models, whereas in the case of *Labor*, "...the larger or more complex classifiers did not diminish the understanding of the decision process, but may have even increased it through providing more steps and including more attributes for each decision step." In an earlier study, Kononenko (1993) found that medical experts rejected rules learned by a decision tree algorithm because they found them to be too short. Instead, they preferred explanations that were derived from a Naïve Bayes classifier, which essentially showed weights for all attributes, structured into confirming and rejecting attributes.

<sup>20</sup> The lack of domain knowledge was hypothesized to account for differences observed in another study by Allahyari and Lavesson (2011), which we discuss in more detail below.

**Rule 1:** if the mushroom has the following properties (simultaneously)

- gill size is *broad* and
- stalk surface above ring is *smooth* and
- mushroom has *one ring* and
- spore print color is *brown*

then the mushroom is edible

**Rule 2:** if the mushroom has the following properties (simultaneously)

- odour is *anise*

then the mushroom is edible

**Which of the rules do you find as more plausible? (required)**

Select one

**?** What is plausibility: seeming reasonable or probable, seeming likely to be true, or able to be believed, possibly true; able to be believed.

**What is the rationale for your choice?**

**?** Please do not include references to answers for previous units. Do not enter text such as 'same' or 'see above'.

**Fig. 6** Example rule pair used in experiments 1–3. For Experiment 3, the description of the rule also contained values of confidence and support, formatted as shown in Fig. 8

To some extent, the results may appear to be inconclusive because the different studies do not clearly discriminate between different aspects of interpretability. Most of the results that report that simpler models are more interpretable refer to syntactic interpretability, whereas, e.g., Allahyari and Lavesson (2011) tackle epistemic interpretability by taking the users' prior knowledge into account. Similarly, the study of Kononenko (1993) has aspects of epistemic interpretability, in that “too short” explanations contradict the experts' experience. pragmatic interpretability of the models has not been explicitly addressed, nor are we aware of any studies that explicitly relate plausibility to model complexity.

#### 4.4 Experiment 1: Are shorter rules more plausible?

Motivated by the somewhat inconclusive evidence in previous works on interpretability and complexity, we set up a crowdsourcing experiment that specifically focuses on the aspect of plausibility. In this and the experiments reported in subsequent sections, the basic experimental setup follows the one discussed in Sect. 3. Here, we only note task-specific aspects.

*Material.* The questionnaires presented pairs of rules as described in Sect. 3.1.2, and asked the participants to give (a) judgment which rule in each pair is more preferred and (b) optionally a textual explanation for the judgment. A sample question is shown in Fig. 6. The judgments were elicited using a drop down box, where the participants could choose from the following five options: “Rule 1 (strong preference)”, “Rule 1 (weak preference)”, “No preference”,

**Table 4** Rule-length experiment statistics

	pairs	judg	qfr	part	Kendall's $\tau$		Spearman's $\rho$	
<i>Traffic</i>	80	408	11	93	0.05	(0.226)	0.06	(0.230)
<i>Quality</i>	36	184	11	41	<b>0.20</b>	(0.002)	<b>0.23</b>	(0.002)
<i>Movies</i>	32	160	5	40	-0.01	(0.837)	-0.02	(0.828)
<i>Mushrooms</i>	10	250	13	84	<b>0.37</b>	(0.000)	<b>0.45</b>	(0.000)
Total	158	1002	11	258				

*pairs* refers to the distinct number of rule pairs, *judg* to the number of trusted judgments, the quiz failure rate *qfr* to the percentage of participants that did not pass the initial quiz as reported by the CrowdFlower dashboard, *part* to the number of trusted distinct workers,  $\tau$  and  $\rho$  to the observed correlation values with *p*-values in parentheses

“Rule 2 (*weak preference*)”, “Rule 2 (*strong preference*)”. As shown in Fig. 6, the definition of plausibility was accessible to participants at all times since it was featured below the drop-down box. As optional input, the workers could provide a textual explanation of their reasoning behind the assigned preference, which we informally evaluated but which is not further considered in the analyses reported in this paper.

*Participants.* The number of judgments per rule pair for this experiment was 5 for the *Traffic*, *Quality*, and *Movies* datasets. The *Mushroom* dataset had only 10 rule pairs, therefore we opted to collect 25 judgments for each rule pair in this dataset.

*Results.* Table 4 summarizes the results of this crowdsourcing experiment. In total, we collected 1002 responses, which is on average 6.3 judgments for each of the 158 rule pairs. On two of the datasets, *Quality* and *Mushroom*, there was a strong, statistically significant *positive* correlation between rule length and the observed plausibility of the rule, i.e., longer rules were preferred. In the other two datasets, *Traffic* and *Movies*, no significant difference could be observed in either way.

In any case, these results show that there is no *negative* correlation between rule length and plausibility. In fact, in two of the four datasets, we even observed a positive correlation, meaning that in these cases, longer rules were preferred.

## 5 The conjunction fallacy

Human-perceived plausibility of a hypothesis has been extensively studied in cognitive science. The best-known cognitive phenomenon related to our focus area of the influence of the number of conditions in a rule on its plausibility is the *conjunctive fallacy*. This fallacy falls into the research program on cognitive biases and heuristics carried out by Amos Tversky and Daniel Kahneman since the 1970s. The outcome of this research program can be succinctly summarized by a quotation from Kahneman’s Nobel Prize lecture at Stockholm University on December 8, 2002:

“... it is safe to assume that similarity is more accessible than probability, that changes are more accessible than absolute values, that averages are more accessible than sums, and that the accessibility of a rule of logic or statistics can be temporarily increased by a reminder.” (Kahneman 2003)

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. 7** The Linda problem (Tversky and Kahneman 1983)

In this section, we will briefly review some aspects of this program, highlighting those that seem to be important for inductive rule learning. For a more thorough review, we refer to Kahneman et al. (1982) and Gilovich et al. (2002), a more recent, very accessible introduction can be found in Kahneman (2011).

### 5.1 The Linda problem

The conjunctive fallacy is in the literature often defined via the “Linda” problem. In this problem, participants are asked whether they consider it more plausible that a person Linda is more likely to be (a) a bank teller or (b) a feminist bank teller (Fig. 7). Tversky and Kahneman (1983) report that based on the provided characteristics of Linda, 85% of the participants indicate (b) as the more probable option. This was essentially confirmed in by various independent studies, even though the actual proportions may vary. In particular, similar results could be observed across multiple settings (hypothetical scenarios, real-life domains), as well as for various kinds of participants (university students, children, experts, as well as statistically sophisticated individuals) (Tentori and Crupi 2012).

However, it is easy to see that the preference for (b) is in conflict with elementary laws of probabilities. Essentially, in this example, participants are asked to compare conditional probabilities  $\Pr(F \wedge B \mid L)$  and  $\Pr(B \mid L)$ , where  $B$  refers to “bank teller”,  $F$  to “active in feminist movement” and  $L$  to the description of Linda. Of course, the probability of a conjunction,  $\Pr(A \wedge B)$ , cannot exceed the probability of its constituents,  $\Pr(A)$  and  $\Pr(B)$  (Tversky and Kahneman 1983). In other words, as it always holds for the Linda problem that  $\Pr(F \wedge B \mid L) \leq \Pr(B \mid L)$ , the preference for alternative  $F \wedge B$  (option (b) in Fig. 7) is a logical fallacy.

### 5.2 The representativeness heuristic

According to Tversky and Kahneman (1983), the results of the conjunctive fallacy experiments manifest that “*a conjunction can be more representative than one of its constituents*”. It is a symptom of a more general phenomenon, namely that people tend to overestimate the probabilities of representative events and underestimate those of less representative ones. The reason is attributed to the application of the *representativeness heuristic*. This heuristic provides humans with means for assessing a probability of an uncertain event. According to the representativeness heuristic, the probability that an object  $A$  belongs to a class  $B$  is



evaluated “by the degree to which *A* is representative of *B*, that is by the degree to which *A* resembles *B*” (Tversky and Kahneman 1974).

This heuristic relates to the tendency to make judgments based on similarity, based on a rule “like goes with like”. According to Gilovich and Savitsky (2002), the representativeness heuristic can be held accountable for a number of widely held false and pseudo-scientific beliefs, including those in astrology or graphology.<sup>21</sup> It can also inhibit valid beliefs that do not meet the requirements of resemblance.

A related phenomenon is that people often tend to misinterpret the meaning of the logical connective ‘and’. Hertwig et al. (2008) hypothesized that the conjunctive fallacy could be caused by “a misunderstanding about conjunction”, i.e., by a different interpretation of ‘probability’ and ‘and’ by the participants than assumed by the experimenters. They discussed that ‘and’ in natural language can express several relationships, including temporal order, causal relationship, and most importantly, can also indicate a collection of sets instead of their intersection. For example, the sentence “He invited friends and colleagues to the party” does not mean that all people at the party were both colleagues and friends. According to Sides et al. (2002), ‘and’ ceases to be ambiguous when it is used to connect propositions rather than categories. The authors give the following example of a sentence that is not prone to misunderstanding: “IBM stock will rise tomorrow and Disney stock will fall tomorrow”. Similar wording of rule learning results may be, despite its verbosity, preferred. We further conjecture that representations that visually express the semantics of ‘and’ such as decision trees may be preferred over rules, which do not provide such visual guidance.

### 5.3 Experiment 2: misunderstanding of ‘and’ in inductively learned rules

Given its omnipresence in rule learning results, it is vital to assess to what degree the ‘and’ connective is misunderstood when rule learning results are interpreted. In order to gauge the effect of the conjunctive fallacy, we carried out a separate set of crowdsourcing tasks. To control for a misunderstanding of ‘and’, the group of workers approached in Experiment 2 additionally received *intersection* test questions which were intended to ensure that all participants understand the *and* conjunction the same way it is defined in the probability calculus. In order to correctly answer these, the respondent had to realize that the antecedent of one of the rules contains mutually exclusive conditions. The correct answer was a weak or strong preference for a rule which did not contain the mutually exclusive conditions.

*Material.* The participants were presented with the same rule pairs as in Experiment 1 (Group 1). The difference between Experiment 1 and Experiment 2 was only one manipulation: instructions in Experiment 2 additionally contained the intersection test questions, not present in Experiment 1. We refer to the participants that received these test questions as Group 2.

*Participants.* Same as for Experiment 1 described earlier. There was one small change for the *Mushroom* dataset, where for economical constraints we collected 15 judgments for each rule pair within Experiment 2, instead of 25 collected in Experiment 1.

*Results.* We state the following proposition: *The effect of higher perceived interpretability of longer rules disappears when it is ensured that participants understand the semantics*

---

<sup>21</sup> Gilovich and Savitsky (2002) give the following example: resemblance of the physical appearance of the sign, such as crab, is related in astrology with personal traits, such as appearing tough on the outside. For graphology, the following example is given: handwriting to the left is used to indicate that the person is holding something back.

**Table 5** Effect of intersection test questions that are meant to ensure that participants understand the logical semantics of ‘and’

Dataset	pairs	Group 1: w/o int. test questions				Group 2: with int. test questions			
		judg	qfr	part	Kendall’s $\tau$	judg	qfr	part	Kendall’s $\tau$
<i>Quality</i>	36	184	11	41	<b>0.20</b> (0.002)	180	31	45	−0.03 (0.624)
<i>Mushroom</i>	10	250	13	84	<b>0.37</b> (0.000)	150	44	54	<b>0.28</b> (0.000)

*pairs* refers to the distinct number of rule pairs, *judg* to the number of trusted judgments, the quiz failure rate *qfr* to the percentage of workers that did not pass the initial quiz as reported by the CrowdFlower dashboard, *part* to the number of trusted distinct workers, and  $\tau$  to the observed correlation values with *p*-values in parentheses

of the ‘and’ conjunction. The corresponding null hypothesis is that the correlation between rule length and plausibility is no longer statistically significantly different from zero for participants that successfully completed the intersection test questions (Group 2). We focus on the analysis on *Mushroom* and *Quality* datasets on which we had initially observed a higher plausibility of longer rules.

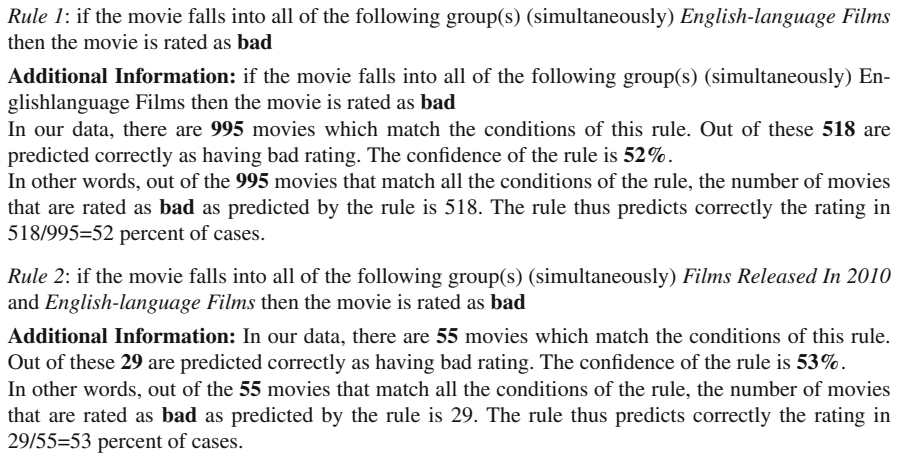
The results presented in Table 5 show that the correlation coefficient is still statistically significantly different from zero for the *Mushroom* dataset with Kendall’s  $\tau$  at 0.28 ( $p < 0.0001$ ), but not for the *Quality* dataset, which has  $\tau$  not different from zero at  $p < 0.05$  (albeit at a much higher variance). This suggests that at least on the *Mushroom* dataset, there are other factors apart from “misunderstanding of and” that cause longer rules to be perceived as more plausible. We will take a look at some possible causes in the following sections.

## 6 Insensitivity to sample size

In the previous sections, we have motivated that rule length is by itself not an indicator for the plausibility of a rule if other factors such as the support and the confidence of the rule are equal. In this and following sections, we will discuss the influence of these and a few alternative factors, partly motivated by results from the psychological literature. The goal is to motivate some directions for future research on the interpretability and plausibility of learned concepts.

### 6.1 Support and confidence

In the terminology used within the scope of cognitive science (Griffin and Tversky 1992), confidence corresponds to the *strength* of the evidence and support to the *weight* of the evidence. Results in cognitive science for the strength and weight of evidence suggest that the weight is systematically undervalued while the strength is overvalued. According to Camerer and Weber (1992), this was, e.g., already mentioned by Keynes (1922), who drew attention to the problem of balancing the likelihood of the judgment and the weight of the evidence in the assessed likelihood. In particular, Tversky and Kahneman (1971) have argued that human analysts are unable to appreciate the reduction of variance and the corresponding increase in reliability of the confidence estimate with increasing values of support. This bias is known as *insensitivity to sample size*, and essentially describes the human tendency to neglect the following two principles: a) more variance is likely to occur in smaller samples, b) larger samples provide less variance and better evidence. Thus, people underestimate the increased benefit of higher robustness of estimates made on a larger sample.



**Fig. 8** Rule pair including the additional information on support and confidence

In the previous experiments, we controlled the rules selected into the pairs, so they mostly had identical or nearly identical confidence and support. Furthermore, the confidence and support values of the shown rules were *not* revealed to the participants during the experiments. However, in real situations, rules in the output of inductive rule learning have varying quality, which is communicated mainly by the values of confidence and support. Given that longer rules can fit the data better, they tend to be higher on confidence and lower on support. This implies that if confronted with two rules of different lengths, where the longer has a higher confidence and the shorter a higher support, the analyst may prefer the longer rule with higher confidence (all other factors equal). These deliberations lead us to the following proposition: *When both confidence and support are explicitly revealed, confidence but not support will positively increase rule plausibility.*

## 6.2 Experiment 3: Is rule confidence perceived as more important than support?

We aim to evaluate the effect of explicitly revealed confidence (strength) and support (weight) on rule preference. In order to gauge the effect of rule quality measures confidence and support, we performed an additional experiment.

*Material.* The participants were presented with rule pairs, like in the previous two experiments. We used only rule pairs from the *Movies* dataset, where the differences in confidence and support between the rules in the pairs were largest. The only difference in the setup between Experiment 1 and Experiment 3 was that participants now also received information about the number of correctly and incorrectly covered instances for each rule, along with the support and confidence values. Figure 8 shows an example of this additional information provided to the participants. Workers that received this extra information are referred to as Group 3.

*Participants.* This setup was the same as for the preceding two experiments.

*Results.* Table 6 shows the correlations of the rule quality measures confidence and support with plausibility. It can be seen that there is a relation to confidence but not to support, even though both were explicitly present in descriptions of rules for Group 3. Thus, our

**Table 6** Kendall's  $\tau$  on the *Movies* dataset with (Group 1) and without (Group 2) additional information about the number of covered good and bad examples

Measure	pairs	Group 1 Without information				Group 3 With information			
		judg	qfr	part	Kendall's $\tau$	judg	qfr	part	Kendall's $\tau$
Support	32	160	5	40	-0.07 (0.402)	160	5	40	-0.08 (0.361)
Confidence	32	160	5	40	0.00 (0.938)	160	5	40	<b>0.24</b> (0.000)

*pairs* refers to the distinct number of rule pairs, *judg* to the number of trusted judgments, the quiz failure rate *qfr* to the percentage of workers that did not pass the initial quiz as reported by the CrowdFlower dashboard, *part* to the number of trusted distinct workers, and  $\rho$  to the observed correlation values with *p*-values in parentheses

result supports the hypothesis that insensitivity to sample size effect is applicable to the interpretation of inductively learned rules. In other words, when both confidence and support are stated, confidence positively affects the preference for a rule, whereas support tends to have no impact.

The results also show that the relationship between revealed rule confidence and plausibility is causal. This follows from confidence not being correlated with plausibility in the original experiment (Group 1 in Fig. 6), which differed only via the absence of the explicitly revealed information about rule quality. While such a conclusion is intuitive, to our knowledge it has not yet been empirically confirmed before.

## 7 Relevance of conditions in rule

An obvious factor that can determine the perceived plausibility of a proposed rule is how relevant it appears to be. Of course, rules that contain more relevant conditions will be considered to be more acceptable. One way of measuring this could be in the strength of the connection between the condition (or conjunction of conditions) with the conclusion. However, in our crowdsourcing experiments, we only showed sets of conditions that are equally relevant in the sense that their conjunction covers about the same number of examples in the shown rules or that the rules have a similar strength of the connection. Nevertheless, the perceived or subjective relevance of a condition may be different for different users.

There are several cognitive biases that can distort the correlation between the relevance of conditions and the judgment of plausibility. One of the most recently discovered is the *weak evidence effect*, according to which evidence in favor of an outcome can actually decrease the probability that a person assigns to it. In an experiment in the area of forensic science reported by Martire et al. (2013), it was shown that participants presented with evidence weakly supporting guilt tended to “invert” the evidence, thereby counterintuitively reducing their belief in the guilt of the accused.

### 7.1 Attribute and literal relevance

In order to analyze the effect of relevance in the rule learning domain, we decided to enrich our input data with two supporting crowdsourcing tasks, which aimed at collecting judgments of attribute and literal relevance.

We kindly ask you to assist us in an experiment that will help researchers understand which properties influence mushroom being considered as poisonous/edible.

Example task follows:

Property: Cap shape

Possible values: bell, conical, convex,  
flat, knobbed, sunken

What is the relevance of the property given above for determining whether a mushroom is edible or poisonous?

Give a judgment on a 10 point scale:

1 = Completely irrelevant  
10 = Very relevant

**Fig. 9** Attribute relevance question for *Mushroom*

*Attribute relevance* Attribute relevance corresponds to human perception of the ability of a specific attribute to predict values of the attribute in rule consequent. For example, in the *Movies* data, the release date of a film may be perceived as less relevant for determining the quality of a film than its language. Attribute relevance also reflects a level of recognition of the explanatory attribute (cf. also Sect. 8), which is a prerequisite to determining the level of association with the target attribute. As an example of a specific attribute that may not be recognized, consider “Sound Mix” for a movie rating problem. This would contrast with attributes such as “Oscar winner” or “year of release”, which are equally well recognized, but clearly associated to a different degree with the target.

*Literal relevance* Literal relevance goes one step further than attribute relevance by measuring human perception of the ability of a specific condition to predict a specific value of the attribute in the rule consequent. It should be noted that we consider the literal relevance to also embed attribute relevance to some extent. For example, the literal (“film released in 2001”) conveys also the attribute (“year of release”). However, in addition to the attribute name, literal also conveys a specific value, which may not be recognized by itself. This again raises the problem of recognition as a prerequisite to association.

## 7.2 Experiment 4: influence of attribute and literal relevance

The experiments were performed similarly as the previous ones using crowdsourcing. Since the relevance experiments did not elicit preferences for rule pairs, there are multiple differences from the setup described earlier. We summarize the experiments in the following, but refer the reader to Kliegr (2017) for additional details.

```

We kindly ask you to assist us in an
experiment that will help researchers
understand which factors can influence
movie ratings.

Example task follows:

Condition: Academy Award Winner or Nominee

The condition listed above will
contribute to a movie being rated as:

    Good (Strong influence)
    Good (Weak influence)
    No influence
    Bad (Weak influence)
    Bad (Strong influence)

Select one option.

```

**Fig. 10** Literal relevance test question for *Movies*

*Material.* The data collected within Experiments 1–3 were enriched with variables denoting the relevance of attributes and literals of the individual rules. Given that in Experiments 1–3 plausibility was elicited for rule pairs, the variables representing relevance were computed as differences of values obtained for the rules in the pair.

Each rule pair was enriched with four<sup>22</sup> variables according to the following pattern: “[Literal|Attribute]Rel[Avg|Max]Δ”. To compute the enrichment variable, the value of the relevance metric for the second rule in the pair ( $r_2$ ) was subtracted from the value for the first rule ( $r_1$ ). For example,

$$LiteralRelAvg\Delta = LiteralRelAvg(r_1) - LiteralRelAvg(r_2), \quad (1)$$

where  $LiteralRelAvg(r_1)$ ,  $LiteralRelAvg(r_2)$  represent the average relevance of literals (conditions) present in the antecedent of rule  $r_1$  ( $r_2$ ) in the pair.

The attribute relevance experiments were prepared for the *Mushroom* and *Traffic* datasets. An example wording of the attribute relevance elicitation task for the *Mushroom* dataset is shown in Fig. 9. An example wording of the literal relevance elicitation task for the *Movies* dataset is shown in Fig. 10. In this case, there was a small difference in setup between the experiments on LOD datasets and the *Mushroom* dataset. The latter task did contain links to Wikipedia for individual literals as these were directly available from the underlying dataset. For the *Mushroom* dataset, no such links were available, and thus these were not included in the task.

<sup>22</sup> We initially also experimented with computing several other enrichment variables not reported here (derived from label length and depth in the taxonomy, using ratios instead of differences, and using minimum in addition to average and maximum). For these variables, we either did not obtain statistically significant results, or the interpretation was not intuitive, therefore, we do not report these additional evaluations here. The interested reader can find some intermediate results in the supplementary material.

**Table 7** Attribute and literal relevance (Group 1, Kendall's  $\tau$ )

Dataset	att	judg	excl	part	Avg		Max	
Attribute relevance								
<i>Traffic</i>	14	35	70	6	0.01	(0.757)	0.00	(0.983)
<i>Mushroom</i>	10	92	66	31	-0.11	(0.018)	<b>0.27</b>	(0.000)
Dataset	lit	judg	excl	part	Avg		Max	
Literal relevance								
<i>Quality</i>	33	165	40	45	<b>0.29</b>	(0.000)	<b>0.31</b>	(0.000)
<i>Movies</i>	30	150	19	40	0.15	(0.012)	<b>0.22</b>	(0.000)
<i>Traffic</i>	58	290	40	75	0.04	(0.311)	0.01	(0.797)
<i>Mushroom</i>	34	170	16	42	<b>-0.19</b>	(0.000)	0.11	(0.037)

Column *att* refers to number of distinct attributes, *lit* to number of distinct literals (attribute-value pairs), *judg* to the number of trusted judgments, *excl* to the percentage of workers that were not trusted on the basis of giving justifications shorter than 11 characters, and *part* to the number of trusted distinct workers

**Results.** Table 7 shows the correlations between plausibility and the added variables representing attribute and literal relevance on the data collected for Group 1 from the previous experiments. The results confirm that literal relevance has a strong correlation with the judgment of the plausibility of a rule. A rule which contained (subjectively) more relevant literals than the second rule in the pair was more likely to be evaluated favorably than rules that do not contain such conditions. This pattern was found valid with varying levels of statistical significance across all evaluation setups in Table 7, except for the average relevance in the smallest *Mushroom* dataset.

Note that the effect is strongest for the maximum relevance, which means that it is not necessary that all the literals are deemed important, but it suffices if a few (or even a single) condition is considered to be relevant. *Traffic* was the only dataset where such effects could not be observed, but this may have to do with the fact that the used attributes (mostly geographic regions) strongly correlate with traffic accidents but do not show a causal relationship. The examination of the relation between the objective relevance of conditions in a rule and their impact on the subjective perception of the rule is an interesting yet challenging area of further study. The perception can be influenced by multiple cognitive phenomena, such as the weak evidence effect.

## 8 Recognition heuristic

The recognition heuristic (Goldstein and Gigerenzer 1999, 2002) is the best-known of the fast and frugal heuristics that have been popularized in several books, such as Gigerenzer et al. (1999, 2011) or Gigerenzer (2015). It essentially states that when you compare two objects according to some criterion that you cannot directly evaluate, and “one of two objects is recognized and the other is not, then infer that the recognized object has the higher value with respect to the criterion.” Note that this is independent of the criterion that should be maximized; it only depends on whether there is an assumed positive correlation with the recognition value of the object. For example, if asked whether Hong Kong or Chongqing is the larger city, people tend to pick Hong Kong because it is better known (at least in the

**Table 8** Correlation of PageRank in the knowledge graph with plausibility (Group 1, Kendall's  $\tau$ )

Dataset	lit	judg	qfr	part	Avg		Max	
<i>Quality</i>	33	165	40	45	0.01	(0.882)	0.07	(0.213)
<i>Movies</i>	30	150	19	40	-0.12	(0.051)	-0.07	(0.275)
<i>Traffic</i>	58	290	40	75	0.03	(0.533)	0.05	(0.195)

Column *lit* refers to number of distinct literals (attribute-value pairs), *judg* to the number of trusted judgments, *qfr* to the percentage of non-trusted workers, and *part* to the number of trusted distinct workers

western hemisphere), even though Chongqing has about four times as many inhabitants. Thus, this criterion may be viewed as being closely associated to relevance, where, in the absence of knowledge about a fact, the city's relevance is estimated by how well it is recognized.

The recognition heuristic can manifest itself as a preference for rules containing a recognized literal or attribute in the antecedent of the rule. Since the odds that a literal will be recognized increase with the length of the rule, it seems plausible that the recognition heuristic generally increases the preference for longer rules. One could argue that for longer rules, the odds of occurrence of an unrecognized literal will also increase. The counterargument is the empirical finding that—under time pressure—analysts assign recognized objects a higher value than to unrecognized objects. This also happens in situations when recognition is a poor cue (Pachur and Hertwig 2006).

### 8.1 Experiment 5: modeling recognition heuristic using PageRank

In an attempt to measure representativeness, we resort to measuring the centrality of a concept using its PageRank (Page et al. 1999) in a knowledge graph. In three of our datasets, the literals correspond to Wikipedia articles, which allowed us to use PageRank computed from the Wikipedia connection graph for these literals. Similarly, as for the previous experiment, each rule pair was enriched with two additional variables corresponding to the difference in the average and maximum PageRank associated with literals in the rules in the pair. We refer the reader to Kliegr (2017) for additional details regarding the experimental setup.

Table 8 shows the correlations between plausibility and the difference in PageRank as a proxy for the recognition heuristic. While we have not obtained statistically strong correlation in the datasets, for two of the datasets (*Quality* and *Traffic*), the direction of the correlation is according to the expectation: plausibility rises with increased recognition. More research to establish the degree of actual recognition and PageRank values is thus needed. Nevertheless, to our knowledge, this is the first experiment that attempted to use PageRank to model recognition.

## 9 Semantic coherence

Thagard (1989) has noted the importance of coherence for explanatory power. This concept is closely related to epistemic interpretability. Note, however, that it is not only important that the explanation is coherent with existing background knowledge, but the explanatory factors should also be coherent with each other, as well as with the concept that should be explained. Thagard (1989) writes that “*a hypothesis coheres with propositions that it explains, or that explain it, or that participate with it in explaining other propositions, or that offer analogous explanations.*”



```

Unemployment = low :- area > 6720,
                    population > 607430.
Unemployment = high :- latitude <= 44.1281,
                    longitude <= 6.3333, longitude > 1.8397.

```

(a) good discriminative rules, highly coherent

```

Unemployment = high :- latitude <= 44.189,
                    population <= 635272
Unemployment = high :- longitude > 1.550,
                    population > 282277

```

(b) good discriminative rules, lowly coherent

**Fig. 11** Example rules for unemployment in different French regions

In previous work (Paulheim 2012a), we conducted experiments with various statistical datasets enriched with Linked Open Data, one being the already mentioned *Quality of Living* dataset, another one denoting the corruption perceptions index (CPI)<sup>23</sup> in different countries worldwide. For each of those, we created rules and had them rated in a user study.

From that experiment, we experienced that many people tend to trust rules more if there is a high semantic coherence between the conditions in the rule. For example, a rule stating the the quality of living in a city is high if it is a *European capital of culture* and is the headquarter of *many book publishers* would be accepted since both conditions refer to cultural topics, whereas a rule involving *European capital of culture* and *many airlines founded* in that city would be considered to be less plausible.

Figure 11 depicts a set of results obtained on an unemployment statistic for French departments, enriched with data from DBpedia (Ristoski and Paulheim 2013). There are highly coherent rules combining attributes such as latitude and longitude, or population and area, as well as lowly coherent rules, combining geographic and demographic indicators. Interestingly, all those combinations perform a similar split of the dataset, i.e., into the continental and overseas departments of France.

At first glance, semantic coherence and discriminative power of a rule look like a contradiction, since semantically related attributes may also correlate: as in the example above, attributes describing the cultural life in a city can be assumed to correlate more strongly than, say, cultural and economic indicators. Hence, it is likely that a rule learner, without any further modifications, will produce semantically incoherent rules at a higher likelihood than semantically coherent ones.

However, in Gabriel et al. (2014), we have shown that it is possible to modify rule learners in a way so that they produce more coherent rules. To that end, attribute labels are linked to a semantic resource such as WordNet (Fellbaum 1998), and for each pair of attributes, we measure the distance in that semantic network. In the first place, this provides us with a measure for semantic coherence within a rule. Next, we can explicitly use that heuristic in the rule learner and combine it with traditional heuristics that are used for adding conditions to a rule. Thereby, a rule learner can be modified to produce rules that are semantically coherent.

The most interesting finding of the above work was that semantically coherent rules could be learned without significantly sacrificing the accuracy of the overall rule-based model. This is possible in cases with lots of attributes that a rule learner can exploit for achieving a

<sup>23</sup> <https://www.transparency.org/research/cpi/overview>.

similar split of the dataset. In the above example with the French departments, any combination of latitude, longitude, population and area can be used to discriminate continental and overseas departments; therefore, the rule learner can pick a combination that has both a high discriminative power and a high coherence.

## 10 Structure

Another factor which, in our opinion, contributes strongly to the interpretability of a rule-based model is its internal logical structure. Rule learning algorithms typically provide flat lists that directly relate the input to the output. Consider, e.g., the extreme case of learning a parity concept, which checks whether an odd or an even number of  $r$  relevant attributes (out of a possibly higher total number of attributes) are set to *true*. Fig. 12a shows a flat rule-based representation of the target concept for  $r = 5$ , which requires  $2^{r-1} = 16$  rules, whereas a structured representation, which introduces three auxiliary predicates (`parity2345`, `parity345`, and `parity45`) is much more concise using only  $2 \cdot (r - 1) = 8$  rules (Fig. 12b). We argue that the parsimonious structure of the latter is much easier to comprehend because it uses only a linear number of rules, and slowly builds up the complex target concept `parity` from the smaller subconcepts `parity2345`, `parity345`, and `parity45`. This is in line with the criticism of Hüllermeier (2015), who argued that the flat structure of fuzzy rules is one of the main limitations of current fuzzy rule learning systems.

However, we are not aware of psychological work that supports this hypothesis. The results of a small empirical validation were recently reported by Schmid et al. (2017), who performed a user study in which the participants were shown differently structured elementary theories from logic programming, such as definitions for `grandfather`, `great-grandfather`, or `ancestor`, and it was observed how quickly queries about a certain ancestry tree could be answered using these predicates. Among other things, the authors posed and partially confirmed the hypothesis that logical programs are more comprehensible if they are structured in a way that leads to a compression in length. In our opinion, further work is needed in order to see whether compression is indeed the determining factor here. It also seems natural to assume that an important prerequisite for structured theories to be more comprehensible is that the intermediate concepts are by themselves meaningful to the user. Interestingly, this was not confirmed in the experiments by Schmid et al. (2017), where the so-called “public” setting, in which all predicates had meaningful names, did not lead to consistently lower answer times than the “private” setting, in which the predicates did not have meaningful names. They also could not confirm the hypothesis that it furthered comprehensibility when their participants were explicitly encouraged to think about meaningful names for intermediate concepts.

Although there are machine learning systems that can tackle simple problems like the family domain, there is no system that is powerful enough to learn deeply structured logic theories for realistic problems, on which we could rely for experimentally testing this hypothesis. In machine learning, this line of work has been known as *constructive induction* (Matheus 1989) or *predicate invention* (Stahl 1996), but surprisingly, it has not received much attention since the classical works in inductive logic programming in the 1980s and 1990s. One approach is to use a wrapper to scan for regularly co-occurring patterns in rules and use them to define new intermediate concepts which allow to compress the original theory (Wnek and Michalski 1994; Pfahringer 1994, 1995). Alternatively, one can directly invoke so-called predicate invention operators during the learning process, as, e.g., in Duce (Muggleton 1987), which operates in propositional logic, and its successor systems in first-order logic (Muggleton

```

parity :-      x1,      x2,      x3,      x4, not x5.
parity :-      x1,      x2, not x3, not x4, not x5.
parity :-      x1, not x2,      x3, not x4, not x5.
parity :-      x1, not x2, not x3,      x4, not x5.
parity :- not x1,      x2, not x3,      x4, not x5.
parity :- not x1,      x2,      x3, not x4, not x5.
parity :- not x1, not x2,      x3,      x4, not x5.
parity :- not x1, not x2, not x2, not x4, not x5.
parity :-      x1,      x2,      x3, not x4,      x5.
parity :-      x1,      x2, not x3, not x4,      x5.
parity :-      x1, not x2,      x3,      x4,      x5.
parity :- not x1,      x2,      x3,      x4,      x5.
parity :- not x1, not x2, not x3,      x4,      x5.
parity :- not x1, not x2,      x3, not x4,      x5.
parity :- not x1,      x2, not x3, not x4,      x5.
parity :-      x1, not x2, not x2, not x4,      x5.

```

(a) flat unstructured rule set

```

parity45 :-      x4,      x5.
parity45 :- not x4, not x5.

parity345 :-      x3, not parity45.
parity345 :- not x3,      parity45.

parity2345 :-      x2, not parity345.
parity2345 :- not x2,      parity345.

parity :-      x1, not parity2345.
parity :- not x1,      parity2345.

```

(b) deep structured rule base with three invented predicates

**Fig. 12** Unstructured and structured rule sets for the parity concept

and Buntine 1988; Kijirikul et al. 1992; Kok and Domingos 2007). One of the few recent works in this area is by Muggleton et al. (2015), who introduced a technique that employs user-provided meta rules for proposing new predicates.

None of these works performed a systematic evaluation of the generated structured theories from the point of view of interpretability. Systems like MOBAL (Morik et al. 1993), which not only tried to learn theories from data but also provided functionalities for reformulating and restructuring the knowledge base (Sommer 1996), have not received much attention in recent years. We believe that providing functionalities and support for learning structured knowledge bases is crucial for the acceptance of learned models in complex domains. In a way, the recent success of deep neural networks needs to be carried over to the learning of deep logical structures. Recent work on so-called sum-product nets, which combine deep learning with graphical models and generate new concepts in their latent variables (Peharz et al. 2017), may be viewed as a step into this direction.

## 11 Conclusion

The main goal of this paper was to motivate that interpretability of rules is an important topic that has received far too little serious attention in the literature. Its main contribution lies in highlighting that plausibility is an important aspect of interpretability, which, to our knowledge, has not been investigated before. In particular, we observed that even rules that have the same predictive quality in terms of conventional measures such as support and confidence, and will thus be considered as equally good explanations by conventional rule learning algorithms, may be perceived with different degrees of plausibility.

More concretely, we reported on five experiments conducted in order to gain a first insight into plausibility of rule learning results. Users were confronted with pairs of learned rules with approximately the same discriminative power (as measured by conventional heuristics such as support and confidence) and were asked to indicate which one seemed more plausible. The experiments were performed in four domains, which were selected so that participants can be expected to be able to comprehend the given explanations (rules), but not to reliably judge their validity without obtaining additional information. In this way, users were guided to give an intuitive assessment of the plausibility of the provided explanation.

Experiment 1 explored the hypothesis whether the Occam's razor principle holds for the plausibility of rules, by investigating whether people consider shorter rules to be more plausible than longer rules. The results obtained for four different domains indicated that this might not be the case; in fact, we observed statistically significant preference for longer rules on two datasets. In Experiment 2, we found support for the hypothesis that the elevated preference for longer rules is partly due to the misunderstanding of 'and' that connects conditions in the presented rules: people erroneously find rules with more conditions as more general. In Experiment 3, we focused on another ingredient of rules: the values of confidence and support metrics. The results suggest that when both confidence and support are stated, confidence positively affects plausibility and support is largely ignored. This confirms a prediction following from previous psychological research studying the *insensitivity to sample size* effect. As a precursor to a follow-up study focusing on the weak evidence effect, Experiment 4 evaluated the relation between perceived plausibility and strength of conditions in the rule antecedent. The results indicate that rule plausibility is affected already if a single condition is considered to be relevant. *Recognition* is a powerful principle underlying many human reasoning patterns and biases. In Experiment 5, we attempted to use PageRank computed from Wikipedia graph as a proxy for how well a given condition is recognized. The results, albeit statistically insignificant, suggest the expected pattern of a positive correlation between recognition and plausibility. This experiment is predominantly interesting from the methodological perspective, as it offers a possible approach to approximation of recognition of rule conditions.

We acknowledge several limitations of the presented experiments. In particular, some of the results might be influenced by the specific domains of the datasets involved. For some of the experiments (Experiment 4), the collected number of judgments was also rather small affecting their statistical significance. Another limitation is the absence of expert users as we relied solely on judgments elicited with crowdsourcing. Lastly, we lacked counsel of a psychologist skilled in designing and evaluating user experiments. Overall, we suggest our experimental results should be replicated on other domains also addressing the other limitations noted above.

In our view, a research program that aims at a thorough investigation of interpretability in machine learning needs to resort to results in the psychological literature, in particular to cognitive biases and fallacies. We summarized some of these hypotheses, such as the

conjunctive fallacy, and started to investigate to what extent these can serve as explanations for human preferences between different learned hypotheses. There are numerous other cognitive effects that can demonstrate how people assess rule plausibility, some of which are briefly listed in Appendix 11 and discussed more extensively in Kliegr et al. (2018). Clearly, more work along these lines is needed.

Moreover, it needs to be considered how cognitive biases can be incorporated into machine learning algorithms. Unlike loss functions, which can be evaluated on data, it seems necessary that interpretability is evaluated in user studies. Thus, we need to establish appropriate evaluation procedures for interpretability and develop appropriate heuristic surrogate functions that can be quickly evaluated and be optimized in learning algorithms.

Finally, in our work, we have largely ignored the issue of background knowledge by picking domains in which we assumed that our participants have a basic knowledge that allows them to judge the plausibility of rules. However, justifiability, i.e., whether a model is in line with existing background knowledge, is an important prerequisite for plausibility. The work of Martens and Baesens (2010) is pioneering in that they try to formalize this notion in the context of domain knowledge. Based on this, another promising research direction is infusing semantic metadata into the learning process and exploiting it for enforcing the output of rules that are likely to be accepted more by the end user.

**Acknowledgements** Open access funding provided by Johannes Kepler University Linz. Much of the work reported in this paper was conducted at TU Darmstadt, Germany, which we gratefully acknowledge. We would like to thank Frederik Janssen and Julius Stecher for providing us with their code, Eyke Hüllermeier, Frank Jäkel, Niklas Lavesson, Nada Lavrač, Ute Schmid, and Kai-Ming Ting for interesting discussions and pointers to related work, and Jilles Vreeken for pointing us to Munroe (2013). We are also grateful for the insightful comments of the anonymous reviewers, which helped us considerably to focus our paper, and provided us with many additional pointers to relevant works in the literature. TK was supported by the Faculty of Informatics and Statistics, University of Economics, Prague through long-term institutional support of research activities and by grant IGA 33/2018. Part of the initial research leading to this article was published in TK's Ph.D. thesis at Queen Mary University of London. Support by Deutscher Akademischer Austauschdienst (DAAD) is also gratefully acknowledged.

**Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

## Appendix: A brief overview of relevant cognitive heuristics and fallacies

In this appendix, we provide a list of cognitive phenomena that can be important for interpretation of rule learning results. However, we neither claim completeness, nor can we provide more than a very short summary of each phenomenon. A more extensive treatment can be found in Kliegr et al. (2018).

The list is divided into three categories. The first two cover cognitive biases (also called illusions) that are included in a recent authoritative review by Pohl (2017). The first category, *Thinking*, covers those related to thinking processes. These require the person to apply a certain rule (such as the Bayes theorem). Since many people do not know this rule, they have to apply it intuitively, which can result in errors. The second category, *Judgment*, covers biases used by people when they are asked to rate some property of a given object (such as a plausibility of a rule). Note that Pohl (2017) also defined a third category, *Memory*, which we do not consider as directly relevant to our problem. Instead, we introduce “*Other*” category into which we put cognitive phenomena that were not explicitly systematized by

Pohl (2017), although many of the phenomena listed under it clearly belong to one of the established categories.

### Thinking

- *Base rate neglect* (Kahneman and Tversky 1973; Bar-Hillel 1980). Insensitivity to the prior probability of the outcome, violating the principles of probabilistic reasoning, especially Bayes' theorem.
- *Confirmation bias and positive test strategy* (Nickerson 1998). Seeking or interpretation of evidence so that it conforms to existing beliefs, expectations, or a hypothesis in hand.
- *Conjunction fallacy and representativeness heuristic* (Tversky and Kahneman 1983). Conjunction fallacy occurs when a person assumes that a specific condition is more probable than a single general condition in case the specific condition seems as more representative of the problem at hand.

### Judgment

- *Availability heuristic* (Tversky and Kahneman 1973). The easier it is to recall a piece of information, the greater the importance of the information.
- *Effect of difficulty* (Griffin and Tversky 1992). If it is difficult to tell which one of two mutually exclusive alternative hypotheses is better because both are nearly equally probable, people will grossly overestimate the confidence associated with their choice. This effect is also sometimes referred to as *overconfidence effect* (Pohl 2017).
- *Mere-exposure effect* (Zajonc 1968). A repeated encounter of a hypothesis results in increased preference.

### Other

- *Ambiguity aversion* (Ellsberg 1961). People tend to favour options for which the probability of a favourable outcome is known over options where the probability of favourable outcome is unknown. Some evidence suggests that ambiguity aversion has a genetic basis (Chew et al. 2012).
- *Averaging heuristic* (Fantino et al. 1997). The joint probability of two events is estimated as an average of probabilities of the component events. This fallacy corresponds to believing that  $P(A, B) = \frac{P(A)+P(B)}{2}$  instead of  $P(A, B) = P(A) * P(B)$ .
- *Confusion of the inverse* (Plous 1993). Conditional probability is equivocated with its inverse. This fallacy corresponds to believing that  $P(A|B) = P(B|A)$ .
- *Context and trade-off contrast* (Tversky and Simonson 1993). The tendency to prefer alternative  $x$  over alternative  $y$  is influenced by the context – other available alternatives.
- *Disjunction fallacy* (Bar-Hillel and Neter 1993). People tend to think that it is more likely for an object to belong to a more characteristic subgroup than to its supergroup.
- *Information bias* (Baron et al. 1988). People tend to believe that more information the better, even if the extra information is irrelevant for their decision.
- *Insensitivity to sample size* (Tversky and Kahneman 1974). Neglect of the following two principles: a) more variance is likely to occur in smaller samples, b) larger samples provide less variance and better evidence.
- *Recognition heuristic* (Goldstein and Gigerenzer 1999). If one of two objects is recognized and the other is not, then infer that the recognized object has a higher value with respect to the criterion.
- *Negativity bias* (Kanouse and Hanson Jr 1987). People weigh negative aspects of an object more heavily than positive ones.

- *Primacy effect* (Thorndike 1927). This effect can be characterized by words of Edward Thorndike (1874–1949), one of the founders of modern educational psychology, as follows: “other things being equal the association first formed will prevail” (Thorndike 1927).
- *Reiteration effect* (Hasher et al. 1977). Frequency of occurrence is a criterion used to establish the validity of a statement.
- *Unit bias* (Geier et al. 2006). People tend to give equal weight to each condition at the expense of detailed scrutiny of its actual weight.
- *Weak evidence effect* (Fernbach et al. 2011). Presenting weak but supportive evidence makes people less confident in predicting a particular outcome than presenting no evidence at all.

While this list is certainly not exhaustive, it is long enough to illustrate that interpretability is a very complex research challenge that cannot be met in passing but needs serious attention in our research programs.

## References

- Agrawal, R., Imielinski, T., & Swami, A. N. (1993). Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data (SIGMOD-93)* (pp. 207–216). Washington, DC.
- Allahyari, H., & Lavesson, N. (2011). User-oriented assessment of classification model understandability. In A. Kofod-Petersen, F. Heintz, & H. Langseth (Eds.), *Proceedings of the 11th Scandinavian conference on artificial intelligence (SCAI-11)* (pp. 11–19). Trondheim: IOS Press.
- Alonso, J. M., Castiello, C., & Mencar, C. (2015). Interpretability of fuzzy systems: Current research trends and prospects. In J. Kacprzyk & W. Pedrycz (Eds.), *Springer handbook of computational intelligence* (pp. 219–237). Berlin: Springer.
- Andrews, R., Diederich, J., & Tickle, A. B. (1995). Survey and critique of techniques for extracting rules from trained artificial neural networks. *Knowledge-Based Systems*, 8(6), 373–389.
- Bar-Hillel, M. (1980). The base-rate fallacy in probability judgments. *Acta Psychologica*, 44(3), 211–233.
- Bar-Hillel, M., & Neter, E. (1993). How alike is it versus how likely is it: A disjunction fallacy in probability judgments. *Journal of Personality and Social Psychology*, 65(6), 1119–1131.
- Baron, J., Beattie, J., & Hershey, J. C. (1988). Heuristics and biases in diagnostic reasoning: II. Congruence, information, and certainty. *Organizational Behavior and Human Decision Processes*, 42(1), 88–110.
- Bensusan, H. (1998). God doesn’t always shave with Occam’s Razor—Learning when and how to prune. In Nédellec, C. & Rouveirol, C. (Eds.), *Proceedings of the 10th European conference on machine learning (ECML-98)* (pp. 119–124).
- Besold, T. R., d’Avila Garcez, A. S., Stenning, K., van der Torre, L. W. N., & van Lambalgen, M. (2017). Reasoning in non-probabilistic uncertainty: Logic programming and neural-symbolic computing as examples. *Minds and Machines*, 27(1), 37–77.
- Bibal, A. & Frénay, B. (2016). Interpretability of machine learning models and representations: An introduction. In *Proceedings of the 24th European symposium on artificial neural networks (ESANN)* (pp. 77–82).
- Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4), 77–84.
- Blumer, A., Ehrenfeucht, A., Haussler, D., & Warmuth, M. K. (1987). Occam’s razor. *Information Processing Letters*, 24, 377–380.
- Bringsjord, S. (2011). Psychometric artificial intelligence. *Journal of Experimental and Theoretical Artificial Intelligence*, 23(3), 271–277.
- Camerer, C., & Weber, M. (1992). Recent developments in modeling preferences: Uncertainty and ambiguity. *Journal of Risk and Uncertainty*, 5(4), 325–370.
- Cano, A., Zafra, A., & Ventura, S. (2013). An interpretable classification rule mining algorithm. *Information Sciences*, 240, 1–20.
- Chaney, A. J., & Blei, D. M. (2012). Visualizing topic models. In *Proceedings of the 6th international conference on weblogs and social media (ICWSM-12)*. Palo Alto: AAAI Press.
- Chew, S. H., Ebstein, R. P., & Zhong, S. (2012). Ambiguity aversion and familiarity bias: Evidence from behavioral and gene association studies. *Journal of Risk and Uncertainty*, 44(1), 1–18.

- Clark, W. A., & Avery, K. L. (1976). The effects of data aggregation in statistical analysis. *Geographical Analysis*, 8(4), 428–438.
- Cohen, W. W. (1995). Fast effective rule induction. In A. Prieditis & S. Russell (Eds.), *Proceedings of the 12th international conference on machine learning (ML-95)* (pp. 115–123). Lake Tahoe, CA: Morgan Kaufmann.
- Craven, M., & Shavlik, J. W. (1997). Using neural networks for data mining. *Future Generation Computing Systems*, 13(2–3), 211–229.
- Crump, M. J., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon’s mechanical Turk as a tool for experimental behavioral research. *PLoS One*, 8(3), e57410.
- Dempster, A. P. (1967). Upper and lower probabilities induced by a multivalued mapping. *The Annals of Mathematical Statistics*, 38(2), 325–339.
- Dhurandhar, A. (2018). How interpretable are you? a framework for quantifying interpretability. In E. Hüllermeier, H. Kestler, & A. Wilhelm (Eds.), *Book of Abstracts of the European Conference on Data Analysis (ECDA-18)* (pp. 58–59). Germany: Paderborn.
- Dhurandhar, A., Iyengar, V., Luss, R., & Shanmugam, K. (2017). *TIP: Typifying the interpretability of procedures*. arXiv preprint [arXiv:1706.02952](https://arxiv.org/abs/1706.02952).
- Domingos, P. (1999). The role of Occam’s Razor in knowledge discovery. *Data Mining and Knowledge Discovery*, 3(4), 409–425.
- Dua, D. & Karra Taniskidou, E. (2017). *UCI machine learning repository*. University of California, School of Information and Computer Science. <http://archive.ics.uci.edu/ml>.
- Duivesteyn, W., Feelders, A., & Knobbe, A. J. (2016). Exceptional model mining—Supervised descriptive local pattern mining with complex target concepts. *Data Mining and Knowledge Discovery*, 30(1), 47–98.
- Elkan, C. (2001). The foundations of cost-sensitive learning. In *Proceedings of the 17th international joint conference on artificial intelligence (IJCAI-01)* (pp. 973–978).
- Ellsberg, D. (1961). Risk, ambiguity, and the savage axioms. *The Quarterly Journal of Economics*, 75(4), 643–669.
- Fantino, E., Kulik, J., Stolarz-Fantino, S., & Wright, W. (1997). The conjunction fallacy: A test of averaging hypotheses. *Psychonomic Bulletin & Review*, 4(1), 96–101.
- Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM*, 39(11), 27–34.
- Fellbaum, C. (Ed.). (1998). *WordNet: An electronic lexical database*. Cambridge, MA: MIT Press.
- Fernbach, P. M., Darlow, A., & Sloman, S. A. (2011). When good evidence goes bad: The weak evidence effect in judgment and decision-making. *Cognition*, 119(3), 459–467.
- Freitas, A. A. (2013). Comprehensible classification models: A position paper. *SIGKDD Explorations*, 15(1), 1–10.
- Fürnkranz, J. (2005). From local to global patterns: Evaluation issues in rule learning algorithms. In K. Morik, J.-F. Boulicaut, & A. Siebes (Eds.), *Local pattern detection* (pp. 20–38). Berlin: Springer.
- Fürnkranz, J., & Flach, P. A. (2005). ROC ‘n’ rule learning—Towards a better understanding of covering algorithms. *Machine Learning*, 58(1), 39–77.
- Fürnkranz, J., & Kliegr, T. (2015). A brief overview of rule learning. In N. Bassiliades, G. Gottlob, F. Sadri, A. Paschke, & D. Roman (Eds.), *Proceedings of the 9th international symposium on rule technologies: Foundations, tools, and applications (RuleML-15)* (pp. 54–69). Berlin: Springer.
- Fürnkranz, J., & Knobbe, A. J. (2010). Guest editorial: Global modeling using local patterns. *Data Mining and Knowledge Discovery*, 21(1), 1–8.
- Fürnkranz, J., Gamberger, D., & Lavrač, N. (2012). *Foundations of Rule Learning*. Berlin: Springer. ISBN 978-3-540-75196-0.
- Furr, R. M., & Bacharach, V. R. (2008). *Psychometrics: An introduction*. Thousand Oaks, CA: Sage.
- Gabriel, A., Paulheim, H., & Janssen, F. (2014). Learning semantically coherent rules. In P. Cellier, T. Charnois, A. Hotho, S. Matwin, M.-F. Moens, & Y. Toussaint (Eds.), *Proceedings of the ECML/PKDD-14 international workshop on interactions between data mining and natural language processing* (pp. 49–63). Nancy: CEUR Workshop Proceedings.
- Gall, R. (2019). Machine learning explainability vs interpretability: Two concepts that could help restore trust in AI. *KDnuggets News*, 19(1). <https://www.kdnuggets.com/2018/12/machine-learning-explainability-interpretability-ai.html>.
- Gamberger, D., & Lavrač, N. (2003). Active subgroup mining: A case study in coronary heart disease risk group detection. *Artificial Intelligence in Medicine*, 28(1), 27–57.
- Ganter, B., & Wille, R. (1999). *Formal concept analysis—Mathematical foundations*. Berlin: Springer.
- Geier, A. B., Rozin, P., & Doros, G. (2006). Unit bias a new heuristic that helps explain the effect of portion size on food intake. *Psychological Science*, 17(6), 521–525.



- Gigerenzer, G. (2015). *Simply rational: Decision making in the real world*. New York: Oxford University Press.
- Gigerenzer, G., Todd, P., & The ABC Group. (1999). *Simple Heuristics that Make us Smart*. Evolution and Cognition Series. Oxford: Oxford University Press.
- Gigerenzer, G., Hertwig, R., & Pachur, T. (Eds.). (2011). *Heuristics: The foundations of adaptive behavior*. New York: Oxford University Press.
- Gillies, M., Fiebrink, R., Tanaka, A., Garcia, J., Bevilacqua, F., Héloir, A., Nunnari, F., Mackay, W. E., Amershi, S., Lee, B., D'Alessandro, N., Tilmanne, J., Kulesza, T., & Caramiaux, B. (2016). Human-Centered Machine Learning. In *Proceedings of the ACM conference on human factors in computing systems (CHI-16)* (pp. 3558–3565). New York: ACM.
- Gilovich, T., & Savitsky, K. (2002). Like goes with like: The role of representativeness in erroneous and pseudo-scientific beliefs. In T. Gilovich, D. Griffin, & D. Kahnemann (Eds.), *Heuristics and biases: The psychology of intuitive judgment chapter 34* (pp. 617–624). Cambridge: Cambridge University Press.
- Gilovich, T., Griffin, D., & Kahnemann, D. (Eds.). (2002). *Heuristics and biases: The psychology of intuitive judgment*. New York: Cambridge University Press.
- Goldstein, D. G. & Gigerenzer, G. (1999). The recognition heuristic: How ignorance makes us smart. In *Simple heuristics that make us smart* (pp. 37–58). Oxford: Oxford University Press.
- Goldstein, D. G., & Gigerenzer, G. (2002). Models of ecological rationality: The recognition heuristic. *Psychological Review*, 109(1), 75–90.
- Griffin, D., & Tversky, A. (1992). The weighing of evidence and the determinants of confidence. *Cognitive Psychology*, 24(3), 411–435.
- Grünwald, P. D. (2007). *The minimum description length principle*. Cambridge: MIT Press. ISBN 9780262072816.
- Hahn, H. (1930). *Überflüssige Wesenheiten: Occams Rasiermesser*. Wien: Veröffentlichungen des Vereines Ernst Mach.
- Hahsler, M., Chelluboina, S., Hornik, K., & Buchta, C. (2011). The arules R-package ecosystem: Analyzing interesting patterns from large transaction data sets. *Journal of Machine Learning Research*, 12, 2021–2025.
- Hasher, L., Goldstein, D., & Toppino, T. (1977). Frequency and the conference of referential validity. *Journal of Verbal Learning and Verbal Behavior*, 16(1), 107–112.
- Hempel, C. G., & Oppenheim, P. (1948). Studies in the logic of explanation. *Philosophy of Science*, 15(2), 135–175.
- Hernández-Orallo, J. (2017). *The measure of all minds—Evaluating natural and artificial intelligence*. Cambridge: Cambridge University Press.
- Hertwig, R., Benz, B., & Krauss, S. (2008). The conjunction fallacy and the many meanings of and. *Cognition*, 108(3), 740–753.
- Hintzman, D. L. (1978). *The psychology of learning and memory*. Dallas: Freeman.
- Hu, J., & Mojsilovic, A. (2007). High-utility pattern mining: A method for discovery of high-utility item sets. *Pattern Recognition*, 40(11), 3317–3324.
- Hu, Z., Ma, X., Liu, Z., Hovy, E. H., & Xing, E.P. (2016). Harnessing deep neural networks with logic rules. In *Proceedings of the 54th annual meeting of the association for computational linguistics (ACL-16), Vol. 1: Long Papers*, Berlin: The Association for Computer Linguistics.
- Hüllermeier, E. (2015). From knowledge-based to data-driven fuzzy modeling—Development, criticism, and alternative directions. *Informatik Spektrum*, 38(6), 500–509.
- Huysmans, J., Dejaeger, K., Mues, C., Vanthienen, J., & Baesens, B. (2011). An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models. *Decision Support Systems*, 51(1), 141–154.
- Jair Escalante, H., Escalera, S., Guyon, I., Baró, X., Güçlütürk, Y., Güçlü, U., et al. (Eds.). (2018). *Explainable and interpretable models in computer vision and machine learning.*, The Springer Series on Challenges in Machine Learning Berlin: Springer.
- Japkowicz, N., & Shah, M. (2011). *Evaluating learning algorithms: A classification perspective*. Cambridge: Cambridge University Press.
- Johnson-Laird, P. N. (1981). Comprehension as the construction of mental models. *Philosophical Transactions of the Royal Society of London Series B, Biological Sciences*, 295, 353–374.
- Kahneman, D. (2003). A perspective on judgment and choice. *American Psychologist*, 58(9), 697–720.
- Kahneman, D. (2011). *Thinking*. Straus and Giroux: Fast and Slow. Farrar. ISBN 9781429969352.
- Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, 80(4), 237–251.
- Kahneman, D., Slovic, P., & Tversky, A. (Eds.). (1982). *Judgment under uncertainty: Heuristics and biases*. New York: Cambridge University Press.

- Kanouse, D. E., & Hanson, L. R., Jr. (1987). Negativity in evaluations. In *Attribution: Perceiving the causes of behavior*. Hillsdale: Lawrence Erlbaum Associates Inc.
- Kemeny, J. G. (1953). The use of simplicity in induction. *The Philosophical Review*, 62(3), 391–408.
- Kendall, M., & Gibbons, J. D. (1990). *Rank correlation methods*. London: Edward Arnold.
- Keynes, J. M. (1922). *A treatise on probability*. London: Macmillan & Co.
- Kijsirikul, B., Numao, M., & Shimura, M. (1992). Discrimination-based constructive induction of logic programs. In *Proceedings of the 10th National Conference on Artificial Intelligence (AAAI-92)* (pp. 44–49).
- Kim, B., Malioutov, D., & Varshney, K. (Eds.). (2016). *Proceedings of the ICML-16 workshop on human interpretability in machine learning (WHI-16)*, New York.
- Kim, B., Malioutov, D. M., Varshney, K. R., & Weller, A. (Eds.). (2017). *Proceedings of the ICML-17 Workshop on Human Interpretability in Machine Learning (WHI-17)*. Sydney: Australia.
- Kim, B., Varshney, K. R., & Weller, A. (Eds.). (2018). *Proceedings of the ICML-18 Workshop on Human Interpretability in Machine Learning (WHI-18)*. Sweden: Stockholm.
- Kleinberg, J. M., Papadimitriou, C. H., & Raghavan, P. (1998). A microeconomic view of data mining. *Data Mining and Knowledge Discovery*, 2(4), 311–324.
- Kliegr, T., Bahník, Š., & Fürnkranz, J. (2018). *A review of possible effects of cognitive biases on interpretation of rule-based machine learning models*. arXiv preprint [arXiv:1804.02969](https://arxiv.org/abs/1804.02969).
- Kliegr, T. (2017). *Effect of cognitive biases on human understanding of rule-based machine learning*. Dissertation Thesis. London: Queen Mary University of London. <https://qmro.qmul.ac.uk/xmlui/handle/123456789/31851>
- Knobbe, A. J., Crémilleux, B., Fürnkranz, J., & Scholz, M. (2008). From local patterns to global models: The LeGo approach to data mining. In A. J. Knobbe (Ed.), *From local patterns to global models: Proceedings of the ECML/PKDD-08 workshop (LeGo-08)* (pp. 1–16), Antwerp, Belgium.
- Kodratoff, Y. (1994). The comprehensibility manifesto. *KDnuggets*, 94(9) (Guest Editor's Introduction, *AI Communications*, 7(2), 83–85).
- Kok, S., & Domingos, P. M. (2007). Statistical predicate invention. In Z. Ghahramani (Ed.), *Proceedings of the 24th international conference on machine learning (ICML-07)* (pp. 433–440), Corvallis: ACM.
- Kononenko, I. (1993). Inductive and Bayesian learning in medical diagnosis. *Applied Artificial Intelligence*, 7, 317–337.
- Kralj Novak, P., Lavrač, N., & Webb, G. I. (2009). Supervised descriptive rule discovery: A unifying survey of contrast set, emerging pattern and subgroup mining. *Journal of Machine Learning Research*, 10, 377–403.
- Lakkaraju, H., Bach, S. H., & Leskovec, J. (2016). Interpretable decision sets: A joint framework for description and prediction. In B. Krishnapuram, M. Shah, A. J. Smola, C. C. Aggarwal, D. Shen, & R. Rastogi (Eds.), *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (KDD-16)* (pp. 1675–1684). San Francisco, CA: ACM.
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444.
- Li, M., & Vitányi, P. (1993). *An introduction to Kolmogorov complexity and its applications*. Berlin: Springer.
- Lincoff, G. H. (1981). *The Audubon society field guide to North American mushrooms*. New York: Knopf.
- Lipton, Z. C. (2016). The mythos of model interpretability. Presented at 2016 ICML Workshop on Human Interpretability in Machine Learning (WHI-16), New York, NY. arXiv preprint [arXiv:1606.03490](https://arxiv.org/abs/1606.03490).
- Markou, M., & Singh, S. (2003a). Novelty detection: A review—Part 1: Statistical approaches. *Signal Processing*, 83(12), 2481–2497.
- Markou, M., & Singh, S. (2003b). Novelty detection: A review—Part 2: Neural network based approaches. *Signal Processing*, 83(12), 2499–2521.
- Martens, D., & Baesens, B., (2010). Building acceptable classification models. In R. Stahlbock, S. F. Crone, & S. Lessmann (Eds.), *Data mining, Vol. 8 of annals of information systems* (pp. 53–74). Berlin: Springer.
- Martens, D., & Provost, F. J. (2014). Explaining data-driven document classifications. *MIS Quarterly*, 38(1), 73–99.
- Martens, D., Vanthienen, J., Verbeke, W., & Baesens, B. (2011). Performance of classification models from a user perspective. *Decision Support Systems*, 51(4), 782–793.
- Martire, K. A., Kemp, R. I., Watkins, I., Sayle, M. A., & Newell, B. R. (2013). The expression and interpretation of uncertain forensic science evidence: verbal equivalence, evidence strength, and the weak evidence effect. *Law and Human Behavior*, 37(3), 197–207.
- Matheus, C.J. (1989). A constructive induction framework. In *Proceedings of the 6th international workshop on machine learning* (pp. 474–475).
- Mayes, G. R. (2001). Theories of explanation. *Internet Encyclopedia of Philosophy*. <https://www.iep.utm.edu/explanat/>.

- Mehta, M., Rissanen, J., & Agrawal, R. (1995). MDL-based decision tree pruning. In U. Fayyad, & Uthurusamy, R. (Eds.), *Proceedings of the 1st international conference on knowledge discovery and data mining (KDD-95)* (pp. 216–221). AAAI Press.
- Michalski, R. S. (1983). A theory and methodology of inductive learning. *Artificial Intelligence*, 20(2), 111–162.
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Weinberger (Eds.) *Advances in neural information processing systems 26 (NIPS)* (pp. 3111–3119).
- Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267(1), 1–38.
- Minnaert, B., Martens, D., Backer, M. D., & Baesens, B. (2015). To tune or not to tune: Rule evaluation for metaheuristic-based sequential covering algorithms. *Data Mining and Knowledge Discovery*, 29(1), 237–272.
- Mitchell, T. M. (1997). *Machine Learning*. New York: McGraw Hill.
- Molnar, C. (2019). *Interpretable machine learning—A guide for making black box models explainable*. <https://christophm.github.io/interpretable-ml-book/>.
- Morik, K., Wrobel, S., Kietz, J.-U., & Emde, W. (1993). *Knowledge acquisition and machine learning—Theory, methods, and applications*. London: Academic Press.
- Muggleton, S. H. (1987). Structuring knowledge by asking questions. In I. Bratko & N. Lavrač (Eds.), *Progress in machine learning* (pp. 218–229). Wilmslow: Sigma Press.
- Muggleton, S. H., & Buntine, W. L. (1988). Machine invention of first-order predicates by inverting resolution. In *Proceedings of the 5th international conference on machine learning (ML-88)* (pp. 339–352).
- Muggleton, S. H., Lin, D., & Tamaddoni-Nezhad, A. (2015). Meta-interpretive learning of higher-order dyadic Datalog: Predicate invention revisited. *Machine Learning*, 100(1), 49–73.
- Muggleton, S. H., Schmid, U., Zeller, C., Tamaddoni-Nezhad, A., & Besold, T. (2018). Ultra-strong machine learning: comprehensibility of programs learned with ILP. *Machine Learning*, 107(7), 1119–1140.
- Müller, K.-R., Vedaldi, A., Hansen, L. K., Samek, W., & Motavon, G. (eds.). (2017). *Proceedings of the NIPS-17 workshop on interpreting, explaining and visualizing deep learning ... now what?* Long Beach.
- Munroe, R. (2013). *Kolmogorov directions*. xkcd.com, A webcomic of romance, sarcasm, math, and language. <https://xkcd.com/1155/>.
- Murphy, P. M., & Pazzani, M. J. (1994). Exploring the decision forest: An empirical investigation of Occam's Razor in decision tree induction. *Journal of Artificial Intelligence Research*, 1, 257–275.
- Needham, S. L., & Dowe, D. L. (2001). Message length as an effective Ockham's Razor in decision tree induction. *Proceedings of the 8th International Workshop on Artificial Intelligence and Statistics (AI+STATS-01)*, Key West, FL (pp. 253–260).
- Newson, R. (2002). Parameters behind “nonparametric” statistics: Kendall's tau, Somers' d and median differences. *The Stata Journal*, 2, 45–64.
- Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, 2(2), 175–220.
- Pachur, T., & Hertwig, R. (2006). On the psychology of the recognition heuristic: Retrieval primacy as a key determinant of its use. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 32(5), 983–1002.
- Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). *The PageRank citation ranking: Bringing order to the web*. Technical Report 1999-66, Stanford InfoLab.
- Paolacci, G., & Chandler, J. (2014). Inside the Turk: Understanding Mechanical Turk as a participant pool. *Current Directions in Psychological Science*, 23(3), 184–188.
- Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making*, 5(5), 411–419.
- Paulheim, H. (2012a). Generating possible interpretations for statistics from linked open data. In *Proceedings of the 9th Extended Semantic Web Conference (ESWC-12)* (pp. 560–574). Berlin: Springer.
- Paulheim, H. (2012b). Nobody wants to live in a cold city where no music has been recorded—Analyzing statistics with explain-a-LOD. In *The Semantic Web: ESWC 2012 Satellite Events*. Heraklion, Crete, Greece. Revised Selected Papers (pp. 560–574). Springer.
- Paulheim, H. & Fürnkranz, J. (2012). Unsupervised generation of data mining features from linked open data. In *Proceedings of the 2nd International Conference on Web Intelligence, Mining and Semantics (WIMS-12)* (pp. 31:1–31:12). ACM.
- Pazzani, M. J. (2000). Knowledge discovery from data? *IEEE Intelligent Systems and their Applications*, 15(2), 10–12.
- Pazzani, M. J., Mani, S., & Shankle, W. R. (2001). Acceptance of rules generated by machine learning among medical experts. *Methods of Information in Medicine*, 40(5), 380–385.

- Peharz, R., Gens, R., Pernkopf, F., & Domingos, P. M. (2017). On the latent variable interpretation in sum-product networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(10), 2030–2044.
- Pfahring, B. (1994). Controlling constructive induction in CiPF: an MDL approach. In P. B. Brazdil (Ed.), *Proceedings of the 7th European Conference on Machine Learning (ECML-94)* (pp. 242–256). Catania: Springer.
- Pfahring, B. (1995). A new MDL measure for robust rule induction (extended abstract). In *Proceedings of the 8th European conference on machine learning (ECML-95)* (pp. 331–334). Berlin: Springer.
- Piatetsky-Shapiro, G. (2018). Will GDPR make machine learning illegal? *KDnuggets*, 18(12). <https://www.kdnuggets.com/2018/03/gdpr-machine-learning-illegal.html>.
- Piltaver, R., Luštrek, M., Gams, M., & Martinčić-Ipšić, S. (2016). What makes classification trees comprehensible? *Expert Systems with Applications*, 62, 333–346.
- Plous, S. (1993). *The Psychology of Judgment and Decision Making*. New York: McGraw-Hill Book Company.
- Pohl, R. (2017). *Cognitive illusions: A handbook on fallacies and biases in thinking, judgement and memory* (2nd ed.). London: Psychology Press.
- Popper, K. R. (1935). *Logik der Forschung: zur Erkenntnistheorie der modernen Naturwissenschaft*. Berlin: Verlag von Julius Springer.
- Popper, K. R. (1959). *The logic of scientific discovery*. London: Hutchinson & Co.
- Post, H. (1960). Simplicity in scientific theories. *The British Journal for the Philosophy of Science*, 11(41), 32–41.
- Quinlan, J. R. (1990). Learning logical definitions from relations. *Machine Learning*, 5, 239–266.
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). “Why should I trust you?”: Explaining the predictions of any classifier. In B. Krishnapuram, M. Shah, A. J. Smola, C. Aggarwal, D. Shen, & R. Rastogi (Eds.) *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-16)* (pp. 1135–1144), San Francisco, CA: ACM.
- Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, 14, 465–471.
- Ristoski, P., de Vries, G. K. D., & Paulheim, H. (2016). A collection of benchmark datasets for systematic evaluations of machine learning on the semantic web. In P. T. Groth, E. Simperl, A. J. G. Gray, M. Sabou, M. Krötzsch, F. Lécué, F. Flöck, & Gil, Y. (Eds.), *Proceedings of the 15th international semantic web conference (ISWC-16), Part II* (pp. 186–194), Kobe, Japan.
- Ristoski, P., & Paulheim, H. (2013). Analyzing statistics with background knowledge from linked open data. In S. Capadisi, F. Cotton, R. Cyganiak, A. Haller, A. Hamilton, & R. Troncy (Eds.), *Proceedings of the 1st international workshop on semantic statistics (SemStats-13)*. CEUR workshop proceedings 1549, Sydney, Australia.
- Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review*, 15(3), 351–337.
- Rothe, S., & Schütze, H. (2016). Word embedding calculus in meaningful ultradense subspaces. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL-15)* (pp. 512–517). Stroudsburg: ACL.
- Schmid, U., Zeller, C., Besold, T., Tamaddoni-Nezhad, A., & Muggleton, S. (2017). How does predicate invention affect human comprehensibility? In J. Cussens, & A. Russo (Eds.), *Proceedings of the 26th international conference on inductive logic programming (ILP-16)* (pp. 52–67). London: Springer.
- Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, 61, 85–117.
- Schmitz, G. P. J., Aldrich, C., & Gouws, F. S. (1999). ANN-DT: An algorithm for extraction of decision trees from artificial neural networks. *IEEE Transactions on Neural Networks*, 10(6), 1392–1401.
- Shafer, G. (1976). *A mathematical theory of evidence*. Princeton: Princeton University Press.
- Sides, A., Osherson, D., Bonini, N., & Viale, R. (2002). On the reality of the conjunction fallacy. *Memory & Cognition*, 30(2), 191–198.
- Smith, E. E., Shoben, E. J., & Rips, L. J. (1974). Structure and process in semantic memory: A featural model for semantic decisions. *Psychological Review*, 1(214–241), 1974.
- Sommer, E. (1996). *Theory Restructuring—A Perspective on Design and Maintenance of Knowledge Based Systems*. Doctoral thesis, Technical University of Dortmund, Germany, volume 171 of DISKI. Infix.
- Stahl, I. (1996). Predicate invention in Inductive Logic Programming. In L. De Raedt (Ed.), *Advances in Inductive Logic Programming* (pp. 34–47). Amsterdam: IOS Press.
- Stecher, J., Janssen, F., & Fürnkranz, J. (September 2014). Separating rule refinement and rule selection heuristics in inductive rule learning. In T. Calders, F. Esposito, E. Hüllermeier, & R. Meo (Eds.) *Proceedings of the European conference on machine learning and principles and practice of knowledge discovery in databases (ECML-PKDD-14), Part 3* (pp. 114–129). Nancy: Springer.
- Stecher, J., Janssen, F., & Fürnkranz, J. (2016). Shorter rules are better, aren't they? In T. Calders, M. Ceci, & D. Malerba (Eds.), *Proceedings of the 19th international conference on discovery science (DS-16)* (pp. 279–294). Berlin: Springer.

- Stumme, G., Taouil, R., Bastide, Y., Pasquier, N., & Lakhal, L. (2002). Computing iceberg concept lattices with Titanic. *Data and Knowledge Engineering*, 42(2), 189–222.
- Tentori, K., & Crupi, V. (2012). On the conjunction fallacy and the meaning of and yet again: A reply to Hertwig, Benz, and Krauss (2008). *Cognition*, 122(2), 123–134.
- Thorndike, E. L. (1927). The influence of primacy. *Journal of Experimental Psychology*, 10(1), 18–29.
- Todorovski, L., Flach, P., & Lavrač, N. (2000). Predictive performance of weighted relative accuracy. In *Proceedings of the 4th European symposium on principles of data mining and knowledge discovery (PKDD-2000)* (pp. 255–264).
- Tosi, A., Vellido, A., & Alvarez, M. (eds.). (2017). *Proceedings of the NIPS-17 workshop on transparent and interpretable machine learning in safety critical environments (TIML-17)*, Long Beach.
- Thagard, P. (1989). Explanatory coherence. *Behavioral and Brain Sciences*, 12(3), 435–467.
- Tran, S. N., & d'Avila Garcez, A. S. (2018). Deep logic networks: Inserting and extracting knowledge from deep belief networks. *IEEE Transactions on Neural Networks and Learning Systems*, 29(2), 246–258.
- Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, 76(2), 105–110.
- Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, 5(2), 207–232.
- Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157), 1124–1131.
- Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological review*, 90(4), 293–315.
- Tversky, A., & Simonson, I. (1993). Context-dependent preferences. *Management Science*, 39(10), 1179–1189.
- Valmarska, A., Lavrač, N., Fürnkranz, J., & Robnik-Sikonja, M. (2017). Refinement and selection heuristics in subgroup discovery and classification rule learning. *Expert Systems with Applications*, 81, 147–162.
- van den Eijkel, G. (1999). Rule induction. In M. Berthold & D. Hand (Eds.), *Intelligent data analysis: An introduction* (pp. 195–216). Berlin: Springer.
- Van Fraassen, B. C. (1977). The pragmatics of explanation. *American Philosophical Quarterly*, 14(2), 143–150.
- Vreeken, J., van Leeuwen, M., & Siebes, A. (2011). Krimp: Mining itemsets that compress. *Data Mining and Knowledge Discovery*, 23(1), 169–214.
- Wallace, C. S., & Boulton, D. M. (1968). An information measure for classification. *Computer Journal*, 11, 185–194.
- Wang, S., Huang, C., Yao, Y., & Chan, A. (2015). Mechanical Turk-based experiment vs laboratory-based experiment: A case study on the comparison of semantic transparency rating data. In *Proceedings of the 29th Pacific Asia conference on language, information and computation (PACLIC-15)*, Shanghai: ACL.
- Wang, T., Rudin, C., Doshi-Velez, F., Liu, Y., Klampfl, E., & MacNeille, P. (2017). A Bayesian framework for learning rule sets for interpretable classification. *Journal of Machine Learning Research*, 18, 70:1–70:37.
- Webb, G. I. (1996). Further experimental evidence against the utility of Occam's razor. *Journal of Artificial Intelligence Research*, 4, 397–417.
- Webb, G. I. (2007). Discovering significant patterns. *Machine Learning*, 68(1), 1–33.
- Weih, S., & Sondhaus, U. M. (2003). Combining mental fit and data fit for classification rule selection. In M. Schwaiger & O. Opitz (Eds.), *Exploratory data analysis in empirical research. Studies in classification, data analysis, and knowledge organization* (pp. 188–203). Berlin: Springer.
- Wille, R. (1982). Restructuring lattice theory: An approach based on hierarchies of concepts. In I. Rival (Ed.), *Ordered Sets* (pp. 445–470). Dordrecht-Boston: Reidel.
- Wilson, A. G., Kim, B., & Herland, W. (Eds.) (2016). *Proceedings of the NIPS-16 workshop on interpretable machine learning for complex systems*. Barcelona, Spain.
- Wnek, J., & Michalski, R. S. (1994). Hypothesis-driven constructive induction in AQ17-HCI: A method and experiments. *Machine Learning*, 14(2), 139–168.
- Zajonc, R. B. (1968). Attitudinal effects of mere exposure. *Journal of Personality and Social Psychology*, 9(2p2), 1–27.
- Zaki, M. J., & Hsiao, C.-J. (2002). CHARM: An efficient algorithm for closed itemset mining. In R. L. Grossman, J. Han, V. Kumar, H. Mannila, & R. Motwani (Eds.), *Proceedings of the 2nd SIAM international conference on data mining (SDM-02)*. Arlington, VA.
- Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In D. Fleet, T. Pajdla, B. Schiele, & T. Tuytelaars (Eds.), *Proceedings of the 13th European conference on computer vision (ECCV-14)* (pp. 818–833). Berlin: Springer.

Zilke, J. R., Loza Mencía, E., & Janssen, F. (2016). DeepRED—Rule extraction from deep neural networks. In T. Calders, M. Ceci, & D. Malerba (Eds.), *Proceedings of the 19th international conference on discovery science (DS-16)* (pp. 457–473). Bari: Springer.

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.