```
[2160|0]  p :- odor = foul.
[1152|0]  p :- gill-color = buff.
[ 256|0]  p :- odor = pungent.
```

(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.
```

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

Fig. 1. Top three rules learned for the class `poisonous` in the *Mushroom* dataset.

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 mined primarily 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.

Consider, e.g., the infamous {`diapers`, `beer`} itemset that is commonly used as an example for a surprising finding in market based analysis. A possible explanation for this finding is that this rule captures the behavior of young family fathers who are sent to shop for their youngster and have to reward themselves with a six-pack. However, if we consider that a young family may not only need beer and diapers, the closed itemset of this particular combination may also include `baby lotion`, `milk`, `porridge`, `bread`, `fruits`, `vegetables`, `cheese`, `sausages`, `soda`, etc. In this extended context, diapers and beer appear to be considerably less surprising. Conversely, an association rule

```
beer :- diapers
```
(1)

with an assumed confidence of 80%, which at first sight appears interesting because of the unexpectedly strong correlation between buying two seemingly unrelated items, becomes considerably less interesting if we learn that 80% of *all* customers buy beer, irrespective of whether they have bought diapers or not. In other words, the association rule 1 is considerably less plausible than the association rule

```
beer:- diapers, baby lotion, milk, porridge, bread,
       fruits, vegetables, cheese, sausages, soda.
```
(2)

even if both rules may have very similar properties in terms of support and confidence.

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. 1 shows the two decision lists that have been found for the UCI *Mushroom* dataset[3] 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. Stecher et al. (2016) and Valmarska et al. (2017) investigated the suitability of such rules for subgroup discovery, with somewhat inconclusive results.

3.3 Conflicting Evidence

There are many plausible 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 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, even in cases where a simpler and a more complex rule covers the same number of examples, shorter rules are not necessarily more understandable. There are a few isolated empirical studies that add to this picture. However, the results on the relation between the size of representation and comprehensibility are limited and conflicting.

Larger Models are Less Comprehensible. Huysmans et al. (2011) were among the first that actually tried to empirically validate the often implicitly made claim that smaller models are more comprehensible. In particular, they related

[3] https://archive.ics.uci.edu/ml/datasets.html.

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 comprehensible, 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 on the domain knowledge the participants of their study had relating to the data used, so that it cannot be ruled out that the obtained result were caused by lack of domain knowledge. A similar study was later conducted by Piltaver et al. (2016), who found a clear relationship between model complexity and comprehensibility in decision trees.

Larger Models are More Comprehensible. A direct evaluation of the perceived understandability 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 estimated domain knowledge of the participants on each of the datasets. On *Labor*, participants were expected to have good domain knowledge but not so for *Contact Lenses*. The study was performed with 100 student subjects 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.

4 The Need for Interpretability Biases

A lot of work in interpretability has focused on the mere syntactic comprehensibility of a concept. For example, Muggleton et al. (2018) provide an operational definition of *comprehensibility*, which essentially captures how quickly a learned concept can be utilized in solving the problems from the same task domain, typically classifying new examples. In Fürnkranz et al. (2018), we have advocated the view that there is more to interpretability than the mere ability to syntactically parse and understand a given concept.

Consider, e.g., Fig. 2, which shows several possible explanations for why a city has a high quality of living, derived by the Explain-a-LOD system, which uses

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

(b) rated lowly by users

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

Linked Open Data as background knowledge for explaining statistics (Paulheim and Fürnkranz, 2012). Clearly, all rules are comprehensible, and can be easily applied in practice. Even though all of them are good discriminators on the provided data and can be equally well applied by an automated system, the first three appear to be more convincing to a human user. However, currently available rule learning systems would not be able to express a preference for the rules in Fig. 2(a) over those in Fig. 2(b). For doing so, one needs to capture not only the comprehensibility of a rule, but also its *plausibility*.

5 Cognitive Biases

In order to work towards interpretability biases for machine learning, it is useful to consider work in psychology on cognitive biases. Tversky and Kahneman (1974) defined a cognitive bias as a " systematic error in judgment and decision-making common to all human beings which can be due to cognitive limitations, motivational factors, and/or adaptations to natural environments. "

The presumably most famous example is the so-called *conjunctive fallacy*, exemplified by the *Linda problem* (cf. Fig. 3). In this problem, subjects 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. 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 by various independent studies, even though the actual proportions may vary. However, of course, hypothesis (a) is more likely to be correct because a conjunction will never cover more cases than each of its constituents. For our purposes, this example reiterates the point that shorter explanations are not necessarily preferred by human subjects, and that a bias for interpretability should take other factors into account.

The conjunctive fallacy has received considerable attention in the psychological literature, and many possible explanations for this and related phenomena have been proposed (cf. Pohl 2017, for a survey). The results are predominantly attributed to the *representative heuristic* (Tversky and Kahneman, 1974),

```
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. 3. The Linda problem (Tversky and Kahneman, 1983).

according to which people tend to confuse probability with similarity, i.e., Linda is more similar to our mental image of a feminist bank teller than to a generic bank teller. Another potentially relevant explanation is given by Hertwig et al. (2008), who hypothesizes that the humans tend to misunderstand conjunctions. They discussed that "and" in natural language can express several relationships, including temporal order, causal relationship, and most importantly, can also indicate a union 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. Moreover, while the conjunctive fallacy is possibly the best-documented result of the representativeness heuristic, there is a number of other cognitive biases and heuristics that can be important for interpretation of rule learning results. A survey of cognitive biases can be found in (Pohl, 2017), and a discussion of their relevance for machine learning in (Kliegr et al., 2018).

6 First Experimental Results

In previous work (Fürnkranz et al., 2018), we have evaluated a selection of cognitive biases in the very specific context of whether minimizing the complexity or length of a rule will also lead to increased interpretability, which is often taken for granted in machine learning research. More concretely, we reported on five crowd-sourcing experiments conducted in order to gain first insights into differences in the 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 respondents 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.

A first experiment 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 showed that this is not the case, in fact we observed statistically significant preference for longer rules on two datasets. In another experiment, 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: some people erroneously find rules with more conditions as more general. A third experiment show that when both confidence and support are explicitly 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 (Tversky and Kahneman, 1971). Other experiments investigated the relevance of attributes and literals used in the conditions of a rule. The results indicated that rule plausibility is affected already if a single condition is considered to be more relevant.[4] In order to investigate the effects of the recognition heuristic (Goldstein and Gigerenzer, 1999), we attempted to use PageRank computed from the Wikipedia knowledge graph as a proxy for how well a given condition is recognized. The results were inconclusive, on one of the datasets we observed plausibility being affected when all conditions in one rule were recognized comparatively more than in the alternative rule.

7 Conclusion

The main goal of this paper was to motivate that interpretability of rules is an important topic, which is more than a simple syntactic readability of the presented models. In particular, we believe that plausibility is an important aspect of interpretability, which, to our knowledge, has received too little attention in the literature. Learners can often find a large variety of rules with the same or similar discriminatory power as measured on hold-out data, but with large difference in their perceived credibility. Machine learning systems need interpretability biases in order to cope with such situations.

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 over different learned hypotheses. 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

[4] Since our experiments were based on subjective comparisons of pairs of rules, a more precise formulation would be, "comparatively more relevant than the most relevant condition in an alternative rule".

interpretability, and develop appropriate heuristic surrogate functions that can be quickly evaluated and optimized in learning algorithms.

Acknowledgements. We would like to thank Frederik Janssen and Julius Stecher for providing us with their code, Eyke Hüllermeier, Frank Jäkel, Niklas Lavesson, Nada Lavrač 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 reviewers of (Fürnkranz et al., 2018), which helped us considerably to focus our paper. TK was supported by grant IGA 33/2018 of the Faculty of Informatics and Statistics, University of Economics, Prague.

References

Allahyari, H., Lavesson, N.: User-oriented assessment of classification model understandability. In: Kofod-Petersen, A., Heintz, F., Langseth, H. (eds.) Proceedings of the 11th Scandinavian Conference on Artificial Intelligence (SCAI-11), pp. 11–19 (2011)

Bensusan, H.: God doesn't always shave with Occam's Razor — learning when and how to prune. In: Nédellec, C., Rouveirol, C. (eds.) ECML 1998. LNCS, vol. 1398, pp. 119–124. Springer, Heidelberg (1998). https://doi.org/10.1007/BFb0026680

Blumer, A., Ehrenfeucht, A., Haussler, D., Warmuth, M.K.: Occam's razor. Inf. Process. Lett. **24**, 377–380 (1987)

Cohen, W.W.: Fast effective rule induction. In: Prieditis, A., Russell, S. (eds.) Proceedings of the 12th International Conference on Machine Learning (ML-95), pp. 115–123. Morgan Kaufmann, Lake Tahoe (1995)

Domingos, P.: The role of Occam's Razor in knowledge discovery. Data Min. Knowl. Discov. **3**(4), 409–425 (1999)

Freitas, A.A.: Comprehensible classification models: a position paper. SIGKDD Explor. **15**(1), 1–10 (2013)

Fürnkranz, J., Flach, P.A.: ROC 'n' rule learning - towards a better understanding of covering algorithms. Mach. Learn. **58**(1), 39–77 (2005)

Fürnkranz, J., Kliegr, T., Paulheim, H.: On cognitive preferences and the interpretability of rule-based models. arXiv preprint arXiv:1803.01316 (2018)

Ganter, B., Wille, R.: Formal Concept Analysis. Springer, Heidelberg (1999). https://doi.org/10.1007/978-3-642-59830-2

Goldstein, D.G., Gigerenzer, G.: The recognition heuristic: how ignorance makes us smart. Simple Heuristics That Make Us Smart, pp. 37–58. Oxford (1999)

Gordon, D.F., DesJardins, M.: Evaluation and selection of biases in machine learning. Mach. Learn. **20**(1–2), 5–22 (1995)

Grünwald, P.D.: The Minimum Description Length Principle. MIT Press, Cambridge (2007)

Hahn, H.: Überflüssige Wesenheiten: Occams Rasiermesser. Veröffentlichungen des Vereines Ernst Mach, Wien (1930)

Hertwig, R., Benz, B., Krauss, S.: The conjunction fallacy and the many meanings of and. Cognition **108**(3), 740–753 (2008)

Huysmans, J., Dejaeger, K., Mues, C., Vanthienen, J., Baesens, B.: An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models. Decis. Support Syst. **51**(1), 141–154 (2011)

Kemeny, J.G.: The use of simplicity in induction. Philos. Rev. **62**(3), 391–408 (1953)

Kliegr, T., Bahník, Š., Fürnkranz, J.: A review of possible effects of cognitive biases on interpretation of rule-based machine learning models. arXiv preprint arXiv:1804.02969 (2018)
Kodratoff, Y.: The comprehensibility manifesto. KDD Nuggets, 94(9) (1994)
Kononenko, I.: Inductive and Bayesian learning in medical diagnosis. Appl. Artif. Intell. **7**, 317–337 (1993)
Li, M., Vitányi, P.: An Introduction to Kolmogorov Complexity and Its Applications. TCS. Springer, New York (2008). https://doi.org/10.1007/978-0-387-49820-1
Mehta, M., Rissanen, J., Agrawal, R.: MDL-based decision tree pruning. In: Fayyad, U., Uthurusamy, R. (eds.) Proceedings of the 1st International Conference on Knowledge Discovery and Data Mining, pp. 216–221. AAAI Press (1995)
Michalski, R.S.: A theory and methodology of inductive learning. Artif. Intell. **20**(2), 111–162 (1983)
Michie, D.: Machine learning in the next five years. In: Proceedings of the 3rd European Working Session on Learning (EWSL-88), pp. 107–122. Pitman (1988)
Mitchell, T.M., The need for biases in learning generalizations. Technical report, Computer Science Department, Rutgers University, New Brunswick (1980)
Mitchell, T.M.: Version spaces: a candidate elimination approach to rule learning. In: Reddy, R. (ed.) Proceedings of the 5th International Joint Conference on Artificial Intelligence (IJCAI-77), pp. 305–310. William Kaufmann (1977)
Mitchell, T.M.: Machine Learning. McGraw Hill, New York (1997)
Muggleton, S.H., Schmid, U., Zeller, C., Tamaddoni-Nezhad, A., Besold, T.: Ultra-strong machine learning: comprehensibility of programs learned with ILP. Mach. Learn. 1–22 (2018)
Munroe, R. Kolmogorov directions. www.xkcd.com, A webcomic of romance, sarcasm, math, and language (2013)
Murphy, P.M., Pazzani, M.J.: Exploring the decision forest: an empirical investigation of Occam's Razor in decision tree induction. J. Artif. Intell. Res. **1**, 257–275 (1994)
Paulheim, H.: Generating possible interpretations for statistics from linked open data. In: Simperl, E., Cimiano, P., Polleres, A., Corcho, O., Presutti, V. (eds.) ESWC 2012. LNCS, vol. 7295, pp. 560–574. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-30284-8_44
Paulheim, H., Fürnkranz, J.: Unsupervised generation of data mining features from linked open data. In: Proceedings of the International Conference on Web Intelligence and Semantics (WIMS'12) (2012)
Piltaver, R., Luštrek, M., Gams, M., Martinčić-Ipšić, S.: What makes classification trees comprehensible? Expert Syst. Appl. **62**, 333–346 (2016)
Pohl, R.: Cognitive Illusions: A Handbook on Fallacies and Biases in Thinking, Judgement and Memory, 2nd edn. Psychology Press, London (2017)
Post, H.: Simplicity in scientific theories. Br. J. Philos. Sci. **11**(41), 32–41 (1960)
Quinlan, J.R.: Learning logical definitions from relations. Mach. Learn. **5**, 239–266 (1990)
Rissanen, J.: Modeling by shortest data description. Automatica **14**, 465–471 (1978)
Schaffer, C.: Overfitting avoidance as bias. Mach. Learn. **10**, 153–178 (1993)
Stecher, J., Janssen, F., Fürnkranz, J.: Separating rule refinement and rule selection heuristics in inductive rule learning. In: Calders, T., Esposito, F., Hüllermeier, E., Meo, R. (eds.) ECML PKDD 2014. LNCS (LNAI), vol. 8726, pp. 114–129. Springer, Heidelberg (2014). https://doi.org/10.1007/978-3-662-44845-8_8
Stecher, J., Janssen, F., Fürnkranz, J.: Shorter rules are better, aren't they? In: Calders, T., Ceci, M., Malerba, D. (eds.) DS 2016. LNCS (LNAI), vol. 9956, pp. 279–294. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46307-0_18

Stumme, G., Taouil, R., Bastide, Y., Pasquier, N., Lakhal, L.: Computing iceberg concept lattices with Titanic. Data Knowl. Eng. **42**(2), 189–222 (2002)
Tversky, A., Kahneman, D.: Belief in the law of small numbers. Psychol. Bull. **76**(2), 105–110 (1971)
Tversky, A., Kahneman, D.: Judgment under uncertainty: heuristics and biases. Science **185**(4157), 1124–1131 (1974)
Tversky, A., Kahneman, D.: Extensional versus intuitive reasoning: the conjunction fallacy in probability judgment. Psychol. Rev. **90**(4), 293–315 (1983)
Valmarska, A., Lavrač, N., Fürnkranz, J., Robnik-Sikonja, M.: Refinement and selection heuristics in subgroup discovery and classification rule learning. Expert Syst. Appl. **81**, 147–162 (2017)
Vreeken, J., van Leeuwen, M., Siebes, A.: Krimp: mining itemsets that compress. Data Min. Knowl. Discov. **23**(1), 169–214 (2011)
Wallace, C.S., Boulton, D.M.: An information measure for classification. Comput. J. **11**, 185–194 (1968)
Webb, G.I.: Further experimental evidence against the utility of Occam's razor. J. Artif. Intell. Res. **4**, 397–417 (1996)
Wille, R.: Restructuring lattice theory: an approach based on hierarchies of concepts. In: Rival, I. (ed.) Ordered Sets, pp. 445–470. Reidel, Dordrecht-Boston (1982)
Zaki, M.J., Hsiao, C.-J.: CHARM: An efficient algorithm for closed itemset mining. In: Grossman, R.L., Han, J., Kumar, V., Mannila, H., Motwani, R. (eds.) Proceedings of the 2nd SIAM International Conference on Data Mining (SDM-02), Arlington (2002)