A Fortiori Case-Based Reasoning: From Theory to Data

Wijnand van Woerkom[1], Davide Grossi[2,3,4], Henry Prakken[1], and Bart Verheij[3]

[1] Department of Information and Computing Sciences, Utrecht University
[2] Bernoulli Institute for Maths, CS and AI, University of Groningen
[3] Institute for Logic, Language and Computation, University of Amsterdam
[4] Amsterdam Center for Law and Economics, University of Amsterdam

Much present-day research is focused on making artificial intelligence (AI) more transparent.[5] This work is partially done in response to mounting concerns that uninterpretable algorithms, so-called 'black box' AI, are making high-impact decisions—such as those with legal, social, or ethical consequences—in an unfair or irresponsible manner. A prominent example of such a system is the proprietary software Correctional Offender Management Profiling for Alternative Sanctions (COMPAS), developed by Northepointe, Inc. for automatic risk assessment of various forms of recidivism, which has seen nationwide use in the United States [1]. Allegations by [1] that COMPAS racially discriminates in its decision-making process have led to a host of follow-up research and discussions. The COMPAS developers have published a response [5], and others have pointed to flaws in the original analysis by ProPublica [16,2]; but as [16] point out, this situation is symptomatic of the larger problem that the use of such black box systems is obstructing independent assessment of bias, regardless of the veracity of the allegations in this particular instance.

Many different kinds of solutions have been proposed, among which those to make AI inherently more transparent [15]; to formulate appropriate regulations [19]; and to monitor the systems and measure bias over time [8]. The line on which the present work builds is that of *post hoc* explainability methods, in which the black-box system is analyzed after it has been trained and little to no access to the way it functions is assumed.

There are in turn many types of post hoc explanation methods, see e.g. [14,20,7]. We will focus on a particular branch originating from the intersection of AI & law, based on *case-based reasoning* (CBR). The idea of a CBR explanation of a decision is to provide an analogy between it and relevant training examples. Proponents of this approach, such as [9], argue that explanations of this form are natural to humans: they are simple, we are well acquainted with reasoning by analogy, and they draw on real evidence in the sense that training examples typically serve as a gold standard that the black box adheres to. Two recent examples of this approach from AI & law are found in the works by [3] and [13].

The method of [13] is based on a formal theory of *precedential constraint*, introduced by [6], which is a formal framework developed to describe the a fortiori reasoning process underlying case law, i.e., to describe the extent to which a body

[5] This extended abstract is based on [18].

of precedents constrains a decision in a new case. The key idea of [13] behind applying this theory is that the training data used by most modern machine learning systems for binary classification—which consists of rows of data for a set of features together with a binary target variable—can be interpreted as the fact situations of legal cases together with their verdicts. On the basis of this 'training examples as cases' interpretation, [13] use the theory of precedential constraint as the theoretical foundation for building a post hoc explanation algorithm. Since the work by [13] several other works have appeared that use this interpretation; such as that by [11,12], for developing post-hoc XAI methods; and by [10], who use the model as a classifier for human-in-the-loop decision support.

The goal of the present work is to investigate the extent to which the 'training examples as cases' interpretation is applicable in practice. We do so in three steps. First, we further develop the theory of precedential constraint by connecting it to order theory and formal (many-sorted) logic. Secondly, we use this connection with logic to write an implementation for computing with the theory of precedential constraint using the Z3 *satisfiability modulo theories* (SMT) solver [4]. For example, this implementation allows us to check whether a case base forces the outcome of a novel fact situation, whether a case base is consistent, and whether a given case is a *landmark*—a case that does not have its outcome determined by the rest of the case base. Thirdly, we use this implementation to analyze various datasets and evaluate the extent to which they obey the a fortiori principle of constraint.

For the data analysis, we first compare the output of our implementation with that of the previous results found by [13]. Then, we instantiate the a fortiori model of precedential constraint on the COMPAS dataset published by [1] and subsequently evaluate various statistics. This real-world data is relevant to concerns driving explainable AI research, and as such it is representative of the situations to which our explanation methods may be applied. For the evaluation, we are interested in the *consistency* percentage, which can be thought of as the degree to which the data obeys the precedent set by other examples. Through this analysis we find that an important role is played by what we shall refer to as landmark cases; those cases that set a new precedent with respect to the other cases. We find that in the case of the COMPAS data, a relatively small number of these landmarks force the decision of almost all other cases. Lastly, we instantiate the model on several datasets recently used by [17]. These datasets are mostly of a synthetic nature and have known ground truth labels expressed by logical formulas. This allows us to thoroughly analyze the degree to which the a fortiori model fits these datasets, and makes full use of the capabilities of our implementation which Z3 affords it.

Acknowledgments. This research was (partially) funded by the Hybrid Intelligence Center, a 10-year programme funded by the Dutch Ministry of Education, Culture and Science through the Netherlands Organisation for Scientific Research, grant number 024.004.022.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Angwin, J., Larson, J., Mattu, S., Kirchner, L.: Machine bias. ProPublica (2016)
2. Barenstein, M.: ProPublica's COMPAS data revisited (2019). https://doi.org/10.48550/arXiv.1906.04711
3. Čyras, K., Satoh, K., Toni, F.: Explanation for case-based reasoning via abstract argumentation. In: Baroni, P., Gordon, T.F., Scheffler, T., Stede, M. (eds.) Computational Models of Argument. Proceedings of COMMA 2016, pp. 243–254. IOS Press (2016). https://doi.org/10.3233/978-1-61499-686-6-243
4. de Moura, L., Bjørner, N.: Z3: An efficient SMT solver. In: Tools and Algorithms for the Construction and Analysis of Systems. pp. 337–340. Springer, Berlin, Heidelberg (2008). https://doi.org/10.1007/978-3-540-78800-3_24
5. Dieterich, W., Mendoza, C., Brennan, T.: COMPAS risk scales: Demonstrating accuracy equity and predictive parity. Research report, Northpointe Inc. Research Department (2016)
6. Horty, J.: Rules and reasons in the theory of precedent. Legal Theory **17**(1), 1–33 (Mar 2011). https://doi.org/10.1017/S1352325211000036
7. Koh, P.W., Liang, P.: Understanding black-box predictions via influence functions. In: Proceedings of the 34th International Conference on Machine Learning. pp. 1885–1894. PMLR (Jul 2017)
8. Kurita, K., Vyas, N., Pareek, A., Black, A.W., Tsvetkov, Y.: Measuring bias in contextualized word representations. In: Proceedings of the First Workshop on Gender Bias in Natural Language Processing. pp. 166–172. Association for Computational Linguistics, Florence, Italy (2019). https://doi.org/10.18653/v1/W19-3823
9. Nugent, C., Cunningham, P.: A case-based explanation system for black-box systems. Artificial Intelligence Review **24**(2), 163–178 (Oct 2005). https://doi.org/10.1007/s10462-005-4609-5
10. Odekerken, D., Bex, F., Prakken, H.: Justification, stability and relevance for case-based reasoning with incomplete focus cases. In: Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law. pp. 177–186. ICAIL '23, Association for Computing Machinery, New York, NY, USA (Sep 2023). https://doi.org/10.1145/3594536.3595136
11. Peters, J.G., Bex, F., Prakken, H.: Justifications derived from inconsistent case bases using authoritativeness. In: 1st International Workshop on Argumentation for eXplainable AI. CEUR Workshop Proceedings, vol. 3209. CEUR, Cardiff, Wales (2022)
12. Peters, J.G., Bex, F., Prakken, H.: Model- and data-agnostic justifications with a fortiori case-based argumentation. In: Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law. pp. 207–216. ICAIL '23, Association for Computing Machinery, New York, NY, USA (2023). https://doi.org/10.1145/3594536.3595164
13. Prakken, H., Ratsma, R.: A top-level model of case-based argumentation for explanation: Formalisation and experiments. Argument & Computation **13**(2), 159–194 (Jan 2022). https://doi.org/10.3233/AAC-210009
14. Ribeiro, M., Singh, S., Guestrin, C.: "Why should I trust you?": Explaining the predictions of any classifier. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations. pp. 97–101. Association for Computational Linguistics, San Diego, California (2016). https://doi.org/10.18653/v1/N16-3020

15. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence **1**(5), 206–215 (May 2019). https://doi.org/10.1038/s42256-019-0048-x
16. Rudin, C., Wang, C., Coker, B.: The age of secrecy and unfairness in recidivism prediction. Harvard Data Science Review **2**(1) (Jan 2020). https://doi.org/10.1162/99608f92.6ed64b30
17. Steging, C., Renooij, S., Verheij, B.: Discovering the rationale of decisions: Towards a method for aligning learning and reasoning. In: Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law. pp. 235–239. ICAIL '21, Association for Computing Machinery, New York, NY, USA (Jul 2021). https://doi.org/10.1145/3462757.3466059
18. van Woerkom, W., Grossi, D., Prakken, H., Verheij, B.: *A Fortiori* case-based reasoning: From theory to data. Journal of Artificial Intelligence Research **81**, 401–441 (Oct 2024)
19. Wachter, S., Mittelstadt, B., Floridi, L.: Why a right to explanation of automated decision-making does not exist in the general data protection regulation. International Data Privacy Law **7**(2), 76–99 (May 2017). https://doi.org/10.1093/idpl/ipx005
20. Wachter, S., Mittelstadt, B., Russell, C.: Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harvard Journal of Law & Technology **31**(2) (2018). https://doi.org/10.2139/ssrn.3063289