

# Rectifying Classifiers

Joint DL+NMR Invited Talk Abstract

Pierre Marquis

*Université d'Artois, CRIL-CNRS, Institut Universitaire de France, F-62300 Lens, France*


## Keywords

Trustworthy AI, belief change, classifier rectification

Dealing with high-risk or safety-critical applications calls for the development of trustworthy AI systems. Beyond prediction, such systems must offer a number of additional facilities, including explanation and verification. The case when the prediction made is deemed wrong by an expert calls for still another operation, called rectification. Rectifying a classifier aims to guarantee that the predictions made by the classifier (once rectified) comply with the expert knowledge. Here, the expert is supposed more reliable than the predictor, but their knowledge is typically incomplete.

Focusing on Boolean classifiers, I will present rectification as a change operation. Following an axiomatic approach, I will give some postulates that must be satisfied by rectification operators. I will show that the family of rectification operators is disjoint from the family of revision operators and from the family of update operators. I will also present a few results about the computation of a rectification operation.


---


 DL 2022: 35th International Workshop on Description Logics, August 7–10, 2022, Haifa, Israel

 [marquis@cril.fr](mailto:marquis@cril.fr) (P. Marquis)

 <https://www.cril.univ-artois.fr/~marquis/> (P. Marquis)

 0000-0002-7979-6608 (P. Marquis)

 © 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)