UNIVERSITÄT LIECHTENSTEIN

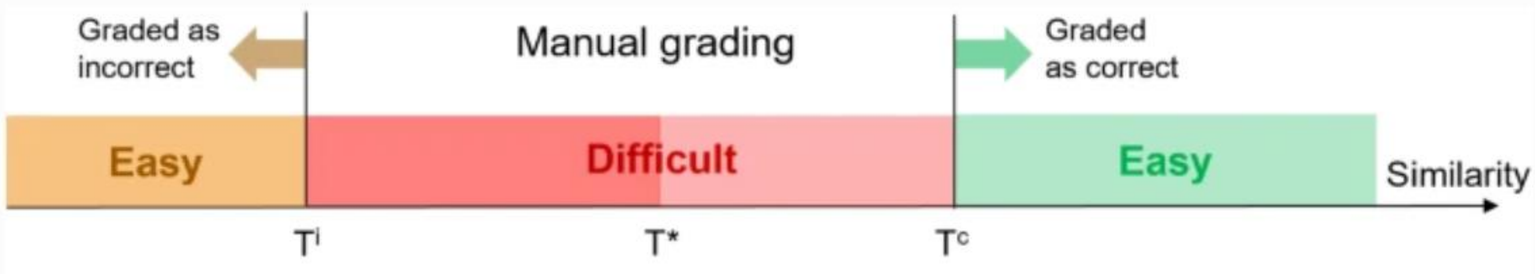
Towards Trustworthy AI: Validating & Explaining AI Models and Decisions

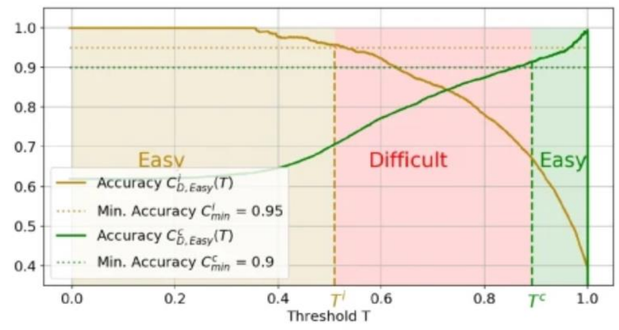
Johannes Schneider

The project covered multiple important questions related to **trustworthy artificial intelligence (AI).** In particular, it looked at validating and explaining AI models and their decisions. One key scenario seeked to validate decisions of an **autograder based on AI, i.e., a system that automatically grades students exams** based on natural language processing (NLP), in order to ensure that the autograder performs consistent across a wide range of exams despite the statistical behavior of an autograder that might allow for very poor grading of exams (occasionally). It should also allow for detailed control of types of errors being made by the system and include humans in the validation procedure. It was executed with the Swiss startup "Classtime". The results were published in a reputable, international journal and showed how a process can look like that leverages autograders based on AI that rely on teachers to validate difficult decisions. Such an approach allows to save millions of hours of teachers worldwide with very little risks of errors in grading. It also gives students the opportunity to practice for exams and get feedback, which would otherwise not be possible.

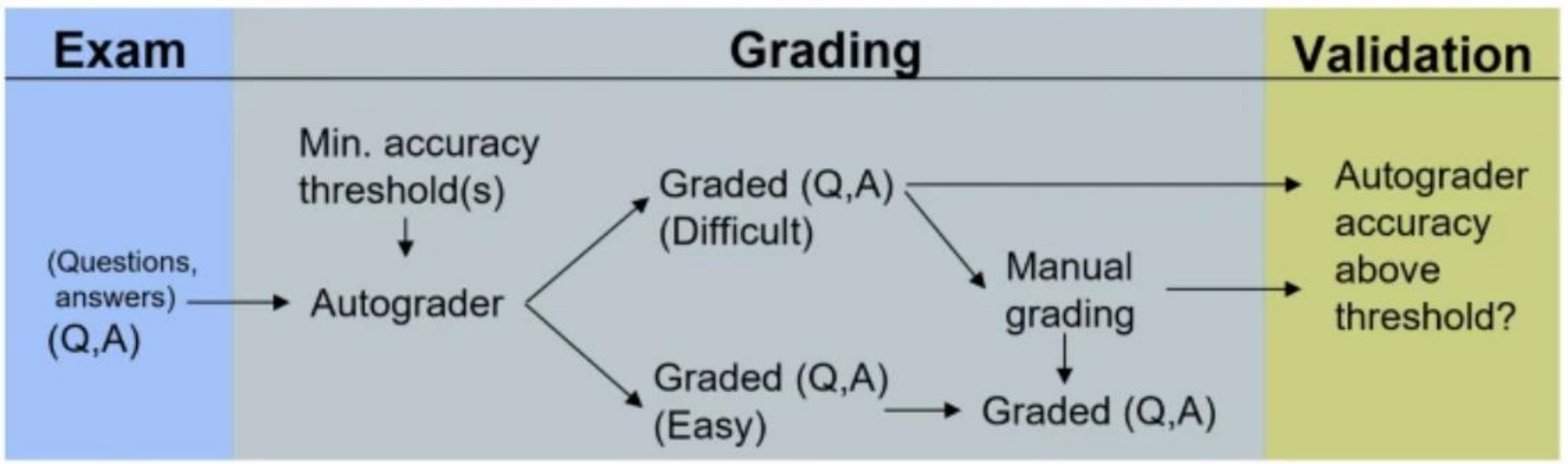
As shown in Fig. 10. The autograder only grades simple questions, while the teacher needs to grade the difficult ones. Difficulty is determined by the autograder, i.e., examples might be simple for a human but difficult for the autograder and vice versa. Teachers can control threshold (Ti,Tc) and thus rely more on the autograder or less for hard questions, relying on the autograder for difficult questions increases the risk of wrong grading (see Figure 10 right panel).

Fig. 10





(c) Accuracy $C^c_{D,Easy}$ and $C^i_{D,Easy}$ depending on the thresholds T^c and T^i.



Overview of approach: Answers are graded, taking into account minimum accuracy thresholds specified by a teacher. Difficult question/answer pairs are (also) manually graded to ensure that accuracy requirements are met. Accuracy on manually and automatically graded answers is compared to validate the performance of the autograder

Paper see:Schneider, J., Richner, R., & Riser, M. (2022). Towards trustworthy autograding of short, multi-lingual, multi-type answers. *International Journal of Artificial Intelligence in Education*, 1-31. https://link.springer.com/article/10.1007/s40593-022-00289-z

Forschungsförderungsfonds