M U N I

Annex No. 10 to the MU Directive on Habilitation Procedures and Professor Appointment Procedures

HABILITATION THESIS REVIEWER'S REPORT

Masaryk University	
Applicant	Mgr. Tomáš Foltýnek, Ph.D.
Habilitation thesis	Information Technologies for Countering Academic Fraud
Reviewer	Assoc. Prof. Daniel Acuña
Reviewer's home unit, institution	University of Colorado at Boulder

The thesis is a commentary document with with ten published articles. The record it presents is strong. The systematic review in ACM Computing Surveys (Article C) is a standard reference for computational plagiarism detection, the definition of plagiarism proposed there reached Council of Europe Recommendation CM/Rec(2022)18, the multilingual test of 15 text-matching systems (Article D) filled a gap that the review itself had identified, and the test of AI-text detectors (Article I) helped move institutional policy away from detector-based enforcement. This is a coherent research program with real effects on practice.

The articles were peer reviewed at their venues, so I have concentrated on the commentary and on three questions that seem to me worth discussing at the defense. I have listed smaller corrections at the end.

Detector failures

Section 2.2 reports that transformer models identified machine-paraphrased text with F1 above 99%, and it adds the important observation that this success was largely caused by atypical word choices introduced by the paraphrasing tool that generated the training data. Chapters 4 and 7 then argue that detection of AI-generated text is fragile and cannot be relied on. These passages sit apart, and the thesis reads as though the applicant had two incompatible results. He does not. They are the same result seen twice.

A detector trained on the output of one generator learns the fingerprints of that generator. Performance is then near perfect in distribution and collapses outside it, which is exactly the pattern Article I documents when paraphrasing and translation are applied. In a review with colleagues (Zhou, Qiu, Liang and Acuna, 2025, IEEE Access) I have argued that detection accuracy in this area is related to the distribution of paraphrase types in the training corpus. Also, the under-representation of some types in widely used datasets is what limits generalization. That framing would let the applicant state a much stronger conclusion than the one currently in Chapter 7. The problem is not that detection is impossible in principle. It is that reported performance is a property of the evaluation corpus rather than of the method, so any single-generator benchmark will overstate what a deployed detector can do.

Reported numbers

The performance figures are reproduced without stating what they measure or against what base rate.

Section 4.1 reports the results of Weber-Wulff et al. (2023), which the applicant co-authored, as accuracy drops for machine-translated text, about 50% for manually edited AI text, and especially for machine-paraphrased text. The published paper reports a drop for translation as well. The figure near 75% is the false negative rate for the paraphrased condition, not an accuracy, I believe. The same section calls that study the first large-scale benchmark of AI-text detectors, which its own related-work tables contradict.

Are the evidence standards are symmetric?

The thesis is demanding of detection technology and generous toward the pedagogical alternatives it recommends.

The abstract states that integrity technologies do not need to be perfect because their presence alone acts as a deterrent. Owens and White (2013), one of the two sources cited, tested the sole use of plagiarism software as a deterrent across ten semesters and found it to be the ineffective baseline against which a writing exercise with feedback produced the reduction. The abstract says the opposite. Youmans (2011), the one randomized test of warning students that their work would be scanned, found no effect and is not cited. Please consider citing.

Section 1.1 also states that honesty pledges and moral reminders dramatically reduce cheating. The reminder result is Mazar, Amir and Ariely (2008), whose Registered Replication Report across 25 labs and 5,786 participants found an effect of d = -0.04, opposite in sign. The thesis notes two sentences earlier that Ariely had serious methodological flaws, and then relies on the findings built on that work. In a thesis about countering academic fraud, this is the passage I would most want revised.

Assessment

The contribution of the work is unusually substantial. The central argument that technologies are limited instruments whose outputs are probabilistic and require human judgment is correct *and* important. The problems are concentrated in the commentary rather than in the published articles, and most are correctable without changing a conclusion.

I recommend that the thesis be accepted.

Reviewer's questions for the habilitation thesis defense (number of questions up to the reviewer)

1. Your paraphrase-detection work reaches F1 above 99% while your AI-detector work concludes that detection is unreliable. Are these the same finding? What does a single-generator benchmark tell us about deployed performance, and how would you design an evaluation that does not overstate it?

2. In Article J, the accuracy of 0.988 and the F1 of 0.937 come from a binomial model assuming independent items and a prior of 0.1. How sensitive are they to each assumption, and what would it take to estimate them from held-out participants?

3. Section 6.2 rejects quiz-based verification because performance tracked working memory rather than authorship. Why does that objection not apply to a cloze test?

4. The abstract states that the presence of integrity technology alone deters, while Section 1.1 and Owens and White (2013) both indicate that a deterrent-only strategy is ineffective. Which do you defend?

5. The evidence for honesty pledges and moral reminders has been retracted or has failed preregistered replication. How would you revise Section 1.1, and what follows for the pedagogical alternative the thesis recommends?

Conclusion

The habilitation thesis entitled *Information Technologies for Countering Academic Fraud* by Mgr. Tomáš Foltýnek, Ph.D. **fulfils** requirements expected of a habilitation thesis in the field of Informatics.

Date: September 8, 2026 Signature: