PROCTORING, PRIVACY AND TRUST: AMERICAN LESSONS FOR EUROPEAN DIGITAL ASSESSMENT

Miriam Solberg
Columbia University, New York, United States

The rapid adoption of remote proctoring during the pandemic left many European universities with tools they had not chosen carefully and cannot easily abandon. This article examines the regulatory and ethical problems that American institutions encountered with remote proctoring and considers what European universities can learn from them. It draws on a review of seventy-two institutional policies, complaints and data protection filings from United States providers, together with interviews with fifteen academic integrity officers. The analysis identifies three failure modes: disproportionate data collection, opacity of algorithmic flagging, and the erosion of trust between students and staff. Each failure mode is traced to a design decision that could have been made differently, and each is contrasted with the European legal framework, which imposes stricter constraints but offers weaker procedural guidance. The article proposes a staged model in which institutions establish a presumption against continuous surveillance, require documented justification for each monitoring measure, and build assessment designs that make surveillance unnecessary. The argument is that assessment integrity is a design property rather than a technological one, and that European universities have an opportunity to demonstrate this.

Contents **Page**

1 Introduction **74**
2 Materials and methods **75**
3 Results **76**
4 Discussion and conclusion **78**

1 Introduction

Between 2020 and 2022, universities in Europe and North America installed remote proctoring software at a speed that bypassed their normal procurement and ethical review processes. The tools recorded students' screens, webcams and sometimes their rooms; they applied machine learning to flag suspicious behaviour and passed the flags to human reviewers; and they produced reports that institutions used in disciplinary proceedings. Five years later, many of those tools remain in use, often without the urgency that justified their adoption and often without any systematic review of whether they work. The pause is therefore an opportunity to ask a question that was not asked at the time: what does proctoring actually do, and at what cost?

The American experience is instructive because it is longer and better documented than the European one. Litigation, regulatory complaints, student protest and scholarly critique have produced a substantial re-

cord of what goes wrong when surveillance is introduced into assessment at scale. That record is not directly transferable: the legal frameworks differ, and the European General Data Protection Regulation imposes constraints that many American vendors were not designed to satisfy. But the failure modes are transferable, because they arise from design choices rather than from legal systems, and design choices cross borders more easily than law does. This article has two aims. The first is analytical: to identify the failure modes of remote proctoring as documented in the American record and to trace each to a specific design decision. The second is practical: to propose an assessment policy that European universities can adopt without either abandoning integrity or importing surveillance. The sections that follow describe the evidence base, present the three failure modes and the model that responds to them, and discuss the conditions under which the model can be implemented.

2 Materials and methods

The study is a structured documentary review combined with expert interviews. The document corpus comprised seventy-two items collected between March and October 2024: institutional policies on remote assessment from forty-one universities, published student complaints and petitions, regulatory filings in three American states, vendor transparency reports where these existed, and peer-reviewed evaluations of proctoring accuracy. Documents were included if they described a specific policy or a specific dispute rather than expressing a general position, a criterion that excluded most commentary and kept the corpus close to practice. Fifteen academic integrity officers were interviewed, drawn from institutions that had either withdrawn a proctoring tool or substantially modified its use. This selection was deliberate. Institutions that had reconsidered their arrangements were more likely to have articulated reasons, and the reasons were the object of the study. Interviews lasted between fifty and ninety minutes, were recorded with consent, and were conducted between April and December 2024.

2.1 Limits of the evidence

The analytical procedure had three steps: coding of documents for decisions, justifications and disputes; identification of failure modes recurring across at least six institutions; tracing each failure mode to a documented design or procurement choice. Three limits shape the conclusions that can be drawn. The corpus is skewed towards institutions that had problems, because institutions with uneventful proctoring do not generate documents. The interview sample is small and

purposive. And the accuracy of proctoring algorithms is documented unevenly: vendors publish little, and independent evaluations are few, so claims about false positive rates rest on a narrow base. These limits mean that the article can describe mechanisms with confidence and estimate frequencies only cautiously.

3 Results

3.1 Disproportionate data collection

The analysis identified three failure modes that recurred across institutions and that were traced to distinct design decisions. They are presented in turn, followed by the pattern that connects them and by the model that responds to it. The first failure mode is the collection of data that is not necessary for the stated purpose. Proctoring tools commonly record the screen, the webcam, the microphone, keystroke timing and sometimes the whole room through a phone camera. Each stream was justified by the vendor as improving detection, and each was accepted by institutions without a separate assessment of necessity. The consequence is that a routine examination generates material of considerable sensitivity: a student's home, family members passing behind them, medical conditions visible in behaviour, and the contents of their computer. Several institutions discovered after adoption that recordings were retained for longer than their own policies allowed, because retention was configured by the vendor rather than by the university.

The proportionality problem is legal as well as ethical. European data protection law requires that processing be limited to what is necessary and proportionate to a specified purpose, and continuous audiovisual recording of a student's room is difficult to defend under that standard when the purpose is to detect unauthorised assistance. American institutions encountered the same argument through a different route, in complaints alleging invasion of privacy and in state legislation restricting biometric and room recording. The interviews suggested that the decisive factor was rarely legal advice; it was the presence of a data protection officer with authority to block procurement before adoption.

3.2 Opacity of algorithmic flagging

The second failure mode concerns flags. Proctoring systems generate suspicion scores from behavioural signals, and those scores are then used, explicitly or implicitly, in disciplinary decisions. The signals themselves are not observable by students, the thresholds are not pub-

lished, and the reasoning behind a particular flag is generally unavailable even to the institution that receives it. This produces two problems at once: students cannot contest a judgement whose basis they cannot see, and institutions cannot demonstrate that a decision was sound. Several integrity officers described receiving a flag report and having no way to distinguish a genuine irregularity from a camera adjustment or a glance at a clock.

3.3 Erosion of trust

The opacity has a distributive consequence that is easy to overlook. False positives cluster among students whose environments are less controlled: those with unstable housing, shared rooms, unreliable bandwidth or disabilities that produce atypical movement. The result is that surveillance falls most heavily on those least able to comply, which is the opposite of what an integrity system should do. Interviews at four institutions described exactly this pattern, and in two of them the pattern had been documented internally without leading to a change of tool. The third failure mode is relational. Students who are monitored continuously describe the examination as an accusation, and staff who rely on flags describe themselves as investigators rather than teachers. The change is difficult to measure and easy to dismiss, but it has practical consequences: students become less willing to report misconduct by others, more likely to contest results, and less likely to accept feedback as developmental. Several interviewees connected the erosion to a rise in formal appeals that were not about marks but about process. Trust, once treated as a soft consideration, turns out to be an operational asset of assessment systems.

The three failure modes are connected by a common assumption: that integrity is produced by observation. Each design decision that produced a failure followed from that assumption, and each alternative that interviewees described followed from its rejection. Institutions that had reconsidered proctoring had not abandoned integrity as a goal; they had relocated it from surveillance to assessment design, using open-book tasks, individualised data sets, oral verification and staged submissions that make unauthorised assistance less useful rather than more detectable.

Mode	Data collected	Principal risk	Practical alternative
Live human invigilation (remote)	Video, audio, screen	Privacy intrusion; scalability	Recorded oral defence in sample
Automated continuous monitoring	Video, audio, screen, keystrokes	Opacity; false positives	Open-book tasks with unique data
Post-hoc integrity analytics	Submission metadata, writing patterns	Attribution errors; appeals	Staged drafts with feedback
No monitoring; design-based integrity	Submission only	Requires assessment redesign	Individualised and applied tasks

Table 1 Proctoring modes, data collected and principal risks. Source: author's review of institutional policies and vendor documentation, 2024.

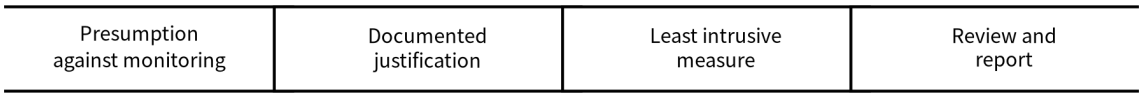

Figure 1 A staged model for justifying monitoring in assessment Source: author's synthesis of interview accounts and policy documents.

4 Discussion and conclusion

The staged model that emerged from the interviews is set out in Figure 1. The first stage is a presumption against monitoring: the default is an assessment design that does not require surveillance, and monitoring must be justified against that default. The second requires a written justification naming the specific risk and why design-based measures are insufficient. The third requires the least intrusive measure capable of addressing that risk. The fourth requires review of whether the measure worked, reported to a body that includes students. The European regulatory environment is stricter than the American one, and that is an advantage rather than a constraint. Institutions that must document necessity and proportionality are forced to make the arguments that their American counterparts were able to skip. The risk is that compliance becomes procedural: a data protection impact assessment completed once, filed, and never revisited, while the tool continues to collect whatever it was configured to collect. The staged model proposed here is intended to prevent that outcome by making justifica-

tion recurrent rather than one-off. Two conditions determine whether the model can be implemented.

The first is institutional: authority to block procurement must sit with someone who has an interest in proportionality, which in practice means the data protection officer rather than the procurement office. The second is academic: assessment redesign requires time that is rarely allocated, and the burden falls on staff who are already overloaded. Universities that have moved away from proctoring have generally done so by funding the redesign explicitly, not by asking staff to absorb it. Without that investment, the presumption against monitoring remains a statement of intent.

The article's limits are those of its evidence: a documentary corpus skewed towards disputes, a small interview sample, and an algorithm literature that is thin on independent evaluation. Future work should examine outcomes rather than perceptions, comparing integrity findings and appeal rates across assessment designs in a controlled way. Until such work exists, the argument rests on a defensible inference: that a system which treats every student as a suspect will produce the behaviour it expects, and that assessment design offers a more reliable route to integrity than observation does.

References

[1] [1] Bretag, T., Harper, R., Burton, M., Ellis, C., Newton, P., Rozenberg, P., Saddiqui, S., and van Haeringen, R. (2019). Contract cheating: a survey of Australian university students. Studies in Higher Education, 44(11), 1837-1856.

[2] [2] Coghlan, S., Miller, T., and Paterson, J. (2021). Good proctor or big brother? Ethics of online exam supervision technologies. Philosophy and Technology, 34(4), 1581-1606.

[3] [3] Dawson, P. (2021). Defending assessment security in a digital world. London: Routledge.

[4] [4] European Data Protection Board. (2021). Guidelines 01/2021 on examples regarding data breach notification. Brussels: EDPB.

[5] [5] Kasprzak, A., and Nixon, L. (2022). Remote proctoring: privacy, equity and the limits of consent. Journal of Academic Ethics, 20(3), 341-359.

[6] [6] Nigam, A., Pasricha, R., Singh, T., and Churi, P. (2021). A systematic review on AI-based proctoring systems. Education and Information Technologies, 26(5), 6421-6445.

[7] [7] Reynders, D. (2020). Remote assessment and the GDPR: necessity and proportionality. European Journal of Law and Technology, 11(3), 1-22.

[8] [8] Slade, C., and Prinsloo, P. (2021). Learning analytics and the algorithmic gaze. British Journal of Educational Technology, 52(4), 1555-1570.

[9] [9] Swauger, S. (2021). Our bodies encoded: algorithmic test proctoring in higher education. Hybrid Pedagogy, 14(2), 1-18.

[10] [10] Vaccaro, A., and Brantmeier, E. (2022). Student trust and assessment integrity in remote examinations. Assessment and Evaluation in Higher Education, 47(8), 1203-1219.