ASCILITE 2025

Future-Focused:

Educating in an Era of Continuous Change

Flagged by design? The intersection of gender, racial and neurodiversity bias in AI proctoring and academic judgment

Mark Gorringe, Karen Williams, Duncan Murray
University of South Australia

While academic misconduct prevention has traditionally focused on student behaviour, limited attention has been paid to the role of educator judgment—particularly under the influence of implicit bias and AI-generated suspicion. As online proctoring software becomes more prevalent, concerns arise regarding systemic disadvantages experienced by specific student groups, particularly female students, students of colour, and neurodiverse learners. This positioning paper explores how AI-driven proctoring technologies, combined with the cognitive demands on academics, may inadvertently amplify reliance on bias and heuristic judgment in academic misconduct decisions. Emerging evidence suggests that certain student groups are disproportionately flagged by proctoring systems and subjected to harsher scrutiny, raising concerns about procedural fairness and equity in online assessments. Rather than reporting empirical findings, this paper outlines a research agenda to investigate how identity-related cues influence both AI flagging and academic judgment. We propose a mixed-method approach—combining meta-analysis with vignette-based quasi-experiments—to critically examine the intersection of bias, surveillance, and academic integrity.

Keywords: online exam proctoring, algorithmic bias, racial/gender bias, neurodiverse students, meta-analysis, case vignette, mixed-method

Introduction

Academic misconduct is not a recent phenomenon (i.e., Whitley & Keith-Spiegel, 2001). Prevention and detection techniques have included prevention approaches relating to course and assessment design, and policing strategies including plagiarism detection tools and terminology detection (Siddhpura & Siddhpura, 2020). However, on the other side of academic integrity equation is the role of the teacher/marker. Human beings are flawed, with often subconscious biases and errors part of our inherent makeup. Greenwald and Banaji (1995, p.8), note that implicit attitudes are defined as "introspectively unidentified (or inaccurately identified) traces of past experience that mediate favourable or unfavourable feeling, thought, or action toward social objects". As noted by Chan (2025, p. 114), stereotypes and bias regarding student identities can:

> *lead to marking bias related to gender (Brennan, 2008; Kiekkas et al., 2016), ethnicity (Bygren, 2020; Lindsey & Crusan, 2011 … Faculty can be affected by implicit bias - "the attitudes or stereotypes that affect our understanding, actions, and decisions in an unconscious manner" (Staats et al., 2017, p. 10).*

Issues regarding student identity and assessment are well documented (see Chan, 2025 for a review). What is less investigated is how student identity might impact academic staff perceptions of which students are cheating. Whilst strategies such as anonymous marking have been considered, such an option is not possible when reviewing breach reports for online exams (i.e., reviewing images of the student in question). This leads to two issues. Firstly, how are students identified as potentially cheating in online proctored examinations? This is often done by proctoring software incorporating AI as a "first check" to determine which students have breached examination conditions. Accordingly, the use of such software may result in 'types' of students that are flagged and presented for review, leading to the generation of particular groups, or stereotypes who become regarded as 'more likely' to cheat. Secondly, academic staff are then required to evaluate, often in limited time given time and pressure constraints faced by modern academics, a potentially large number of breach reports. Fair, (2023, p. 347), identifies the potential for such time pressures to result in academic staff

deferring to AI decision-making in identifying academic misconduct, citing, "Given that faculty face time constraints, these tools could prompt instructors to look for academic fraud disproportionately among one group of students, resulting in a systematically biased group of students being reported".

Research Aim:

We contend that inherent biases in AI proctoring software, combined with the time and pressure constraints faced by the modern academic, create a "perfect storm", unintentionally exacerbating academic reliance on heuristics and biases when making decisions on academic misconduct. We seek to explore the potential for differential disadvantage of specific student groups by considering the following three research questions:

1. Does AI Proctored software result in certain groups being identified as more likely to "cheat"
2. If so, then do preexisting biases among academics exist when determining if academic misconduct has occurred in the cases put before them?
3. Does who (i.e., characteristics of students) influence perceptions of academic misconduct?

Online exam proctoring

Online exam proctoring has become a widespread practice in higher education, particularly since the COVID-19 pandemic accelerated the shift to remote learning (Jiang et al., 2023). These systems typically use a combination of AI-powered surveillance tools such as facial recognition, eye-tracking, and keystroke pattern analysis to identify behaviour considered consistent with cheating (Oravec, 2022). Online proctoring systems like ProctorU, Respondus, PSI and ExamSoft use these AI surveillance tools, in some cases in combination with human invigilators, to detect suspicious behaviour. While these tools promise efficiency and fairness, emerging evidence suggests they may inadvertently reinforce systemic biases (Coglan et al., 2020), given that these systems typically rely on machine learning algorithms trained on datasets that lack demographic diversity. Underlying this is a normative assumption about what constitutes "typical" behaviour during an exam (Broussard, 2023). These assumptions are often implicitly based on Western, neurotypical, and able-bodied standards (Dawson, 2024). Students from different cultural backgrounds may have distinct communication styles or body language that diverge from these expectations. For instance, in some cultures, direct eye contact may be perceived as disrespectful, yet proctoring software may treat lack of eye contact as a sign of dishonesty (Uono & Hietanen, 2015). This paper focusses on three groups that may be notably impacted upon by these systemic biases: female students, students of colour and neurodiverse students

Biases against female students

Forgas (2011) suggested that student appearance may systematically influence evaluative judgments. In their study, identical philosophical essays were assessed more favourably when accompanied by a photograph of a middle-aged, bespectacled man compared to a photograph of a younger female. This highlights the presence of halo effects whereby initial visual impressions bias subsequent academic evaluations. There is also evidence that female students tend to experience heightened stress and anxiety in exams (Chung et al., 2024). This may result in behaviours by female students that result in greater perception of "breaches" of exam conditions (i.e., looking away, head down, eyes darting around etc). Still, the research is equivocal. Butler-Henderson and Crawford's (2020) systematic review of student experiences in online assessments noted that, although one study investigated gender effects and found no significant differences between proctored and non-proctored exams, there is still a lack of large-scale research specifically examining gender in the context of online proctoring. Accordingly potential for gender disparity in this area needs further exploration.

Biases against students of colour

Systemic biases in educational assessment that disadvantage students of colour have been identified. For example, Chowdhury et al. (2020) found that students with Chinese names were less likely to be awarded a

"bare" pass compared to students with traditionally "white" names. AI systemic biases within online proctoring systems can lead to facial recognition failures and disproportionate flagging for "suspicious" behaviour. Buolamwini and Gebru (2018) found that facial analysis algorithms had error rates as high as 34.7% for darker-skinned women, compared to 0.8% for lighter-skinned men. In exam settings, such disparities can lead to unfair accusations of misconduct. Yoder-Himes, et al. (2022) found that students with darker skin tones were significantly more likely to be flagged for instructor review. While no significant differences between male and female students were found, women with the darkest skin tones were flagged the most frequently, suggesting an inherent algorithmic bias operating along intersectional lines. In addition to unfair accusations of misconduct the psychological burden of being surveilled and potentially mischaracterised can significantly impact student performance (Chung et al., 2024). For students of colour, this may compound existing stereotype threats—the fear of confirming negative group stereotypes—which are known to impair test performance (Steele & Aronson, 1995).

Biases against neurodiverse students

Neurodiverse individuals are also disproportionately affected by proctoring systems. These tools often penalise behaviours such as stimming, frequent eye movement, or the need to stand up or fidget—behaviours that might be essential for the student's concentration and self-regulation, but which are flagged as "abnormal" (i.e., Le Cunff et al., 2024). Moreover, strict rules about staying in frame or maintaining continuous eye contact can create high levels of anxiety and reduce performance quality. For neurodiverse students, the lack of flexibility and understanding in proctoring software can turn exams into exclusionary experiences.

The need for investigation

In sum, we have highlighted how the intersection of AI-driven proctoring technologies and the cognitive demands placed on academics may unintentionally reinforce reliance on bias and heuristic judgment in academic misconduct determinations. We outline evidence that certain groups may be disproportionately flagged by proctoring systems and judged more harshly by academic staff, raising serious concerns about procedural fairness and equity in online assessment. This paper serves as a positioning study as, rather than presenting final empirical findings, it sets out a research agenda aimed at investigating how identity-related cues influence both the flagging of students by AI proctoring systems and the judgments made by academic staff. Through the integration of meta-analysis and vignette-based quasi-experimental methods, we propose a comprehensive and empirically grounded approach to critically examine the intersection of bias, surveillance technology, and academic integrity.

Method

To address the research questions, we plan to examine the issue via a mixed-method approach, with RQ1 being investigated via a meta-analysis of relevant literature, and RQ2 and 3 employing a case vignette approach.

RQ 1

For RQ1 we propose a meta-analytic research design to systematically investigate and quantify the presence of bias in proctoring software, specifically in relation to gender, neurodiversity, and race. A meta-analysis is particularly suitable as it allows for the aggregation and critical synthesis of existing empirical findings across diverse studies, enhancing generalisability and power of the conclusions (Thacker, 1988). A comprehensive literature search will be conducted using academic databases including Scopus, Web of Science, PsycINFO, and PubMed. Search terms will include combinations of keywords such as *"online proctoring"*, *"automated invigilation"*, *"remote exams"*, *"algorithmic bias"*, *"racial bias"*, *"gender discrimination"*, *"neurodiverse students"*, *"machine learning fairness"*, and *"educational technology equity"*. Boolean operators (AND/OR) will be used to optimize the sensitivity and specificity of the search. Studies will be included if they met the following criteria:

- Peer-reviewed empirical studies published between 2015 and 2025.
- Focus on online or AI-driven proctoring systems used in educational settings.

- Examination of differential impacts or detection accuracy across gender, racial/ethnic identity, or neurodiverse characteristics (e.g., ADHD, autism spectrum conditions).
- Quantitative studies reporting effect sizes, group comparisons, or statistical measures relevant to bias or disparate outcomes.

Exclusion criteria will include:

- Opinion pieces, commentaries, or non-peer-reviewed sources.
- Studies focused solely on academic integrity without analysis of identity-related impacts.
- Research on non-proctoring educational technologies.

Key data extracted will include: study characteristics, participant demographics, type of proctoring software used, the dimension(s) of identity examined, reported outcomes (e.g., false positive rates, flagging rates), and effect sizes where available. Studies will be coded independently by two of the authors, with discrepancies resolved through consultation with the third author. Where effect sizes are provided a random-effects meta-analysis will be conducted to account for heterogeneity across studies. Subgroup analyses are planned to examine whether the magnitude of bias varied based on identity category (gender, neurodiversity, race), type of proctoring system (AI-based vs. human-monitored), and educational context (e.g., higher education vs. K–12). Publication bias will be assessed through funnel plots and Egger's regression test (Borenstein et al., 2009).

RQ 2/3 -

Participants for this stage of the study will be approximately 340 academic staff experienced in grading online exams invigilated via proctoring software. Participants will be recruited via professional networks, university mailing lists, and teaching and learning forums. Participation will be voluntary with informed consent obtained prior to data collection. Demographic information will be collected to explore any moderating variables in assessment judgments. The study will use a set of standardised case vignettes as per Thyer (2011), simulating academic integrity reports generated by proctoring software. Each vignette will include:

- A brief description of a student's behaviour during an online exam (reflecting one of five types of breaches, ranging in severity):
    1. Administrative breach (e.g., looking away briefly, slight mispositioning of webcam)
    2. Environmental breach (e.g., someone briefly entering the room)
    3. Unverifiable activity (e.g., ambiguous hand movements)
    4. Potential academic misconduct (e.g., use of a phone)
    5. Clear academic misconduct (e.g., accessing course notes or internet resources)
- A still image of a student simulating one taken within an exam using a proctoring platform, will be displayed alongside the vignette. Six identity cues will be embedded across vignettes:
    - A white male student (serving as the "control/neutral" condition)
    - A white female student
    - A male student of colour (e.g., visibly of African, South Asian, or Middle Eastern descent)
    - A female student of colour
    - A male student with visible cues of neurodiversity (e.g., noise-cancelling headphones, fidget device, caption indicating diagnosis such as ASD or ADHD)
    - A female student with visible cues of neurodiversity

All materials will be pre-tested for face validity and neutrality in language and structure, and identities represented consistently across conditions to ensure comparability. Participants will be asked to rate the severity of each breach on a Likert scale comprising six items (rated from 1 = no concern to 5 = serious academic misconduct), and to indicate their willingness on a 5 point Likert scale (1 = not at all willing to 5 – very willing) to pursue one of five actions (e.g., no action, warning, penalty, fail grade, academic misconduct referral). They will also be requested to provide justifications for their decisions.

Participants will complete the study via an online survey platform, being randomly presented with one of 30 vignettes, representing all combinations of student identity (6 levels) × breach type (5 levels). The instructions will emphasise that the proctoring software has flagged the behaviours and that the final decision on whether academic misconduct has occurred is at the lecturer's (i.e., their) discretion, mimicking real-world conditions.

The data will be analysed using ANOVA and mixed-effects models to detect whether student identity systematically influences perceived severity or recommended penalties across different types of breaches.

References

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Wiley. https://doi.org/10.1002/9780470743386

Broussard, M. (2023). *More than a glitch*. MIT Press.

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In Friedler, S. A. & Wilson, C. (Eds.), *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, PMLR 81:77-91 https://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf

Butler-Henderson, K., & Crawford, J. (2020). A systematic review of online examinations: A pedagogical innovation for scalable authentication and integrity. *Computers & Education, 159*, 104024. https://doi.org/10.1016/j.compedu.2020.104024

Chan, L. (2025). Exploring Anonymous Marking to Mitigate Marking Bias: A Self-Study Through Mixed Methods Action Research. *The Canadian Journal of Action Research*, *25*(1), 113–135. https://doi.org/10.33524/cjar.v25i1.685

Chowdhury, S., Klauzner, I., & Slonim, R. (2020). What's in a name? Does racial or gender discrimination in marking exist*? IZA Discussion Paper No. 13890*. Institute of Labor Economics (IZA). https://doi.org/10.2139/ssrn.3734758

Chung, J., Yu, A.S., & Henderson, M. (2024). Inequity in exam experience: Large-scale survey of proctoring and online exam experience by gender, internationality and language. In Cochrane, T., Narayan, V., Bone, E., Deneen, C., Saligari, M., Tregloan, K., Vanderburg, R. (Eds.), Navigating the Terrain: Emerging frontiers in learning spaces, pedagogies, and technologies. *Proceedings ASCILITE 2024. Melbourne (pp. 156-164).* https://doi.org/10.14742/apubs.2024.1229

Coghlan, S., Miller, T., & Paterson, J. (2020). Good proctor or "Big Brother"? AI ethics and online exam supervision technologies. *Ethics and Information Technology, 22*(4), 277–289.

Dawson, P. (2024). Remote proctoring: Understanding the debate. In: Eaton, S.E. (Ed.), *Second Handbook of Academic Integrity* (pp. 1511-1526). Springer International Handbooks of Education. Springer, Cham. https://doi.org/10.1007/978-3-031-54144-5_150

Forgas, J. P. (2011). She just doesn't look like a philosopher…? Affective influences on the halo effect in impression formation. *European Journal of Social Psychology, 41*(7), 812-817 https://doi.org/10.1002/ejsp.842

Jiang, X., Goh, T.-T., & Chen, X., Liu, M., & Yang, B. (2023). Investigating university students' online proctoring acceptance during COVID-19: An extension of the technology acceptance model. *Australasian Journal of Educational Technology*, *39*(2), 47-64. https://doi.org/10.14742/ajet.8121

Le Cunff, A. L., Giampietro, V., & Dommett, E. (2024). Neurodiversity and cognitive load in online learning: A focus group study. *PloS one*, *19*(4), e0301932. https://doi.org/10.1371/journal.pone.0301932

Oravec, J. A. (2022). AI, biometric analysis, and emerging cheating detection systems: The engineering of academic integrity? *Education Policy Analysis Archives, 30*, 5765. https://doi.org/10.14507/epaa.30.5765

Siddhpura, A., & Siddhpura, M. (2020). Plagiarism, contract cheating and other academic misconducts in online engineering education: Analysis, detection and prevention strategies. In *2020 IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE)*, Takamatsu, Japan: 112-119 https://doi.org/10.1109/TALE48869.2020.9368311

Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology, 69*(5), 797–811. https://doi.org/10.1037/0022-3514.69.5.797

Thacker S. B. (1988). Meta-analysis. A quantitative approach to research integration. *JAMA*, *259*(11), 1685–1689. https://doi.org/10.1001/jama.259.11.1685

Thyer, B. A. (2011). *Quasi-experimental research designs: A practical guide*. Oxford University Press.

Uono, S., & Hietanen, J. K. (2015). Eye contact perception in the West and East: a cross-cultural study. *PloS one*, *10*(2), e0118094. https://doi.org/10.1371/journal.pone.0118094

Whitley, Jr., B. E., & Keith-Spiegel, P. (2001). *Academic Dishonesty*. Psychology Press. https://doi.org/10.4324/9781410604279

Yoder-Himes, D., Ross, E., & Shekhar, V. (2022). Racial, skin tone, and sex disparities in automated proctoring software. *Frontiers in Education, 7*, 881449. https://doi.org/10.3389/feduc.2022.881449

Gorringe, M., Williams, K. & Murray, D. (2025). Flagged by design? The intersection of gender, racial and neurodiversity bias in AI proctoring and academic judgment. In Barker, S., Kelly, S., McInnes, R., & Dinmore, S. (Eds.), *Future Focussed. Educating in an era of continuous change*. Proceedings ASCILITE 2025. Adelaide (pp. 425-430). https://doi.org/10.14742/apubs.2025.2687