Dear Reviewers,

Thank you very much for your feedback! We have addressed the areas for improvement mentioned by the reviewers and updated the paper accordingly. Please see the responses (in bold) to the individual reviewers below:

Reviewer ZeLV

> This paper introduces a human-in-the-loop AI cheating ring detection system, showcasing its strength in detecting patterns of cheating behavior and evaluating its performance and fairness. However, it lacks in-depth discussions on the limitations of fairness metrics, potential weaknesses of the system, and unintended consequences such as privacy concerns and impact on different demographic groups. Recommendations include providing more details on limitations and biases, discussing long-term implications, including controlled studies, and addressing the adaptability of the system to new cheating strategies.
>
> Strengths:
>
> One of the key strengths of the paper is the introduction of a human-in-the-loop AI cheating ring detection system, which represents a new paradigm in research. The system is designed to detect patterns of behavior that indicate cheating, such as the impersonation of test-takers by professional cheaters. The authors have successfully illustrated the methodologies used to evaluate the system's performance and fairness, which is crucial given the profound implications on test takers.
>
> The system's evaluation is thorough and the performance of different methods, including keystroke baseline, deep neural networks using keystrokes, and mouse features, is compared, with the deep-keystroke+mouse method achieving the best performance.
>
> The paper also addresses the fairness of the method, aiming to equalize the true negative rate (TNR) across different demographic groups, which is a significant consideration in the development of AI systems.
>
> Weaknesses:
>
> While the paper provides a comprehensive overview of the system's design and evaluation, there are areas where more detail or analysis could strengthen the research. For instance, the paper could benefit from a more in-depth discussion of the limitations of the current fairness metrics and alternative approaches to assess fairness. Additionally, the paper does not explicitly mention any weaknesses of the proposed system, which is a critical aspect of a balanced research review.
>
> The paper could also explore the potential unintended consequences of the system, such as privacy concerns and the impact on different demographic groups, which have been raised in the context of proctoring software. Furthermore, the paper does not address the potential for new forms of cheating that may emerge as AI models become more sophisticated.

Recommendations for Improvement:

The authors should consider providing more details on the system's limitations and potential biases. This includes a deeper analysis of the fairness metrics used and alternative approaches to ensure fairness across demographic groups. Additionally, the paper could benefit from a discussion on the long-term implications of AI proctoring on student privacy and the potential for reinforcing societal biases.

> **[Response] Thank you for your insightful suggestions. We have expanded our discussion to more thoroughly address the system's limitations. Our current fairness metric, TNR, is aimed at equitable treatment of innocent test takers across various groups. However, we acknowledge that this does not ensure uniform detection rates of cheaters across these groups. As a future research direction, we plan to explore additional fairness metrics for a more comprehensive assessment of fairness. Moreover, in the discussion section, we highlighted the needs to adhere to responsible AI standards, which include protecting student privacy and preventing potential societal biases.**

It would also be valuable to include controlled studies that demonstrate the effectiveness of the system in detecting and preventing cheating, as the current literature lacks such evidence. Finally, the authors should explore how the system will adapt to new cheating strategies and the evolving landscape of online assessments.

> **[Response] Thank you for your feedback. We acknowledge the importance of evaluating the system's effectiveness, which we currently track through metrics such as the percentage of test takers flagged. While controlled studies to further evaluate effectiveness are beyond our current project scope, we have identified this as an area for future research. Additionally, in the discussion section, we emphasize the importance of continually adapting the system to new cheating strategies and evolving online assessment environments.**

Reviewer Vdxj

Summary: This paper proposes a machine-learned model to detect cheating by matching keyboard and mouse patterns. They empirically evaluate fairness of their cheating detection algorithm by comparing true negative rates across demographic groups.

Strengths: the problem of detecting cheating rings seems well-motivated. The paper also experimentally looks at the choice of decision thresholds for a low false positive rate and evaluates differences in FPRs across groups. It makes sense to me that the focus would initially be on FPR's, but perhaps the authors can elaborate on the practical implications (as they already do somewhat in the last paragraph of the system evaluation).

Weaknesses:

Given the description of how the data is constructed, it's not clear to me why one would expect differences in FPRs across demographic groups in the evaluation of fairness. For instance, were some demographic groups more heavily represented in the data than others?

> **[Response] Thank you for your comments. The sample was chosen to reflect the diversity of the test taker population. However, it is important to note that within this population, certain demographic groups might be more heavily represented than others. This imbalance in representation might contribute to variations in the FPRs across different groups. We have updated the table to include the ratio of each group of pairs within the dataset. Regarding the fairness evaluation, our current choice of fairness metric (i.e., TNR) is designed to ensure that innocent test takers are treated equitably across different groups. However, it does not inherently guarantee that cheaters are detected with equal probability across these groups. We have expanded the discussion in our paper to address the limitations of using TNR as a sole fairness metric and acknowledge that it is not the only way to assess fairness.**

Also, did the experiments involve selecting different per-group thresholds, or are the reported FNRs under the same decision threshold across all groups?

> **[Response] As for the decision threshold, the same threshold is used for all groups. While we also considered adjusting thresholds for different demographic groups, which might allow us to maintain similar TNRs across groups even when we tolerate a lower TNR, we found it challenging in finding a good threshold selection method. In particular, because a pair of test sessions often involves test takers from different demographic groups across different different demographic attributes, unless we break the symmetry of the predictions, it would be difficult to adjust the thresholds to achieve the desired results across the groups. (Note: since the cheater can impersonate test takers from any demographic group, we cannot limit the pairs to be constructed from test takers who are in the same demographic group.) In addition, having a single threshold also provides additional benefits of avoiding the need to use the demographic attributes as inputs to the system.**

Reviewer 1kNs

The paper introduces a human-in-the-loop cheating detection system based on principles of responsible AI.

The strengths of the work are as follows:

The paper clearly motivates the need for cheating detection in online test-taking.

The proctoring interface is an interesting concept, and there is significant synergy between the work and the data/services provided by online test-taking software.

I appreciate the inclusion of a fairness/ethicality assessment of your system. Not only is "responsible AI" ensuring fairness for students vs. cheaters, but the work takes it a step further to ensure fairness amongst demographics within students vs. cheaters.

The primary weakness of the work is as follows:

The scenario you focus on is a bit unclear to me. Does the professional cheater have control of the periphery, but the test session is still owned by the client? In that case, the location and image would be of the client. Or does the cheater log in on behalf of the client? In that case, the location and image would be of the cheater. My understanding of the paper would be different in these two perspectives -- clarifying this would be helpful.

> **[Response] Thank you for your comments. The proposed system is designed for detecting professional cheaters in both scenarios you described: whether the cheater is only controlling the periphery or logging in on behalf of the client. The core of our detection mechanism hinges on the analysis of unique keystroke and mouse movement patterns. These patterns are key indicators of professional cheating behavior and tend to remain consistent across multiple test sessions, even when conducted by the same cheater for different clients. This consistency in behavior patterns is the primary focus of our detection system, rather than the visual elements of the client's image or location. While similarities in image and location across different clients could indeed serve as additional corroborating evidence of professional cheating, the proposed system focus on behavioral patterns (e.g., keystroke, mouse movements) in order to ensure the system's robustness in scenarios where visual elements like images or location data might be inconsistent or deliberately altered.**

Overall, this work would be a great addition to the workshop.

Best,

Authors of Human-in-the-Loop AI for Cheating Ring Detection