Irish College of Humanities and Applied Sciences

2024

2024 Annual Quality Report ICHAS CASE STUDIES RELATED to Academic Year September 2022 – August 2023

CASE STUDY 3

Title: Identifying unapproved use of Generative AI

Theme: 3. Development and use of Learner Assessment.

Rationale: The most significant Learner Assessment issue during the reporting period pertained to immediate availability of AI technology

Keywords (2-3): Generative AI, Detection Software, Academic Integrity.

Insert Case Study 1 below (in any format - QQI does not prescribe):

Case Description:

Generative AI can be succinctly understood as "computational techniques that are capable of generating seemingly new, meaningful content such as text images or audio from training data" (Feuerriegel et al, 2024, p. 111). Crucially it applies techniques like Reinforcement Learning from Human Feedback (RLHF), that can deliver "nuanced and contextually aware responses across a wide array of conversational topics" (Deng, Zhao and Huang, 2023, p.4770). ChatGPT, one such platform, launched in November 2022; becoming the fastest internet service to reach the one million user milestone at that time, doing so in just five days. Instagram, the previous record holder, had taken two and a half months to reach that milestone while it had taken Netflix three and half years. ChatGPT had reached 100 million active users by January 2023 and recorded 1.3 billion users in September 2023 (Deng, Zhao and Huang, 2023, p.4770). It became immediately apparent that higher education students were among those using the platform. As early as January 2023 a survey of almost 4500 Stanford students revealed 17% had used ChatGPT for Fall Quarter assignments (Allen Cu and Hochman, 2023). That ChatGPT had launched in the latter period of the semester, indicated the rapidity with which it had been adopted and purposed by students.

College Faculty and Management were debriefed by the IT Manager on the significance and possible implications of ChatGPT in early December, 2022. While the full potential of Generative AI, both positive and negative, was appreciated from an early stage in internal discussions, the most immediate task was to consider the implications for the maintenance of academic integrity. There was little that could be done in real terms as approximately 50% of the semester's assignments had already been submitted at that point, restricting the scope for any intervention. Assurances were provided by the College's existing plagiarism detection software provider: "The AI writing indicator that has been added to the Similarity Report will show an overall percentage of the document that may have been AI-generated. We make this determination with 98% confidence based on data that was collected and verified in our AI Innovation Lab" (Turnitin, 2024, n.p.). However, Faculty felt that to use an untested technology and to initiate any form of retrospective enforcement was not feasible. For the College, the shift toward Authentic Assessment, while purely incidental in this context, provided considerable mitigation and it was agreed that any improper use of AI would be considered individually and formatively.

A number of assessment types were identified as vulnerable, most especially, those categorised as terminal assessment. Validatory culture based on reasonable interpretations of best practice, has created a normative expectation to preserve typological spread when designing assessment tasks. In this case, the "project" type of assessment was least affected with the "terminal" type (e.g. essay, literature review, dissertation) most affected by Generative AI. This has since been borne out evidentially (Hanover Research, 2023).

It emerged in first and second marking of first semester assignments that anomalies were occurring in some submissions. A small number of assignments were flagged by AI detector software and were subsequently reviewed by the Moderation Committee. A review of existing policies and procedures was also initiated arising from the Committee's findings. The component that will be explored in this Case Study pertained to the status of detection software within existing policies and procedures.

Case Analysis

Sectoral responsivity to disruptive technology could be generally characterised as a process of measured progressive assimilation. Viewed as a disruptive technology, what was perhaps most unique about Generative AI was not its function or capability, but how the traditional means of response to technological disruption was derailed by the rapidity of its adoption and the coping capacity of obverse technologies. The higher education sector was perhaps uniquely placed in this respect. The far-reaching societal impacts of Generative AI would certainly be far more profound but were not experienced with the same level of immediacy after the initial rollout.

Primarily conceptualised as a socio-technical system, it was clear that the implications of generative AI technology had a wide reach and a sobering balance sheet. It was estimated for example that commercial application of AI could increase global Gross Domestic Product (GDP) by 7% while at the same time resulting in the loss of 300 million jobs; primarily in the knowledge economy (Goldman Sachs, 2023). Within the education sector, the implications similarly inferred equal potential for threat and opportunity. With the launch of ChatGPT coinciding with the first assessment cycle of the new academic year, the negatively disruptive aspects of the service were perhaps more immediately apparent and what received the most attention.

Beyond the implications for academic integrity, other sectoral concerns were mirrored in Faculty discussions. These included flawed and inaccurate data generated by ChatGPT; a lack of regulation internally and externally; increased commercialisation and resultant unequal access to the technology; and the potential for the recycling of cognitive or cultural bias within the source data (UNESCO, 2023). Generative AI's lack of capacity for moral or ethical "thought" is also a matter of controversy and of import in disciplines where professional training, and by implication assessment, is driven by complex ethical and value frameworks. It is also noteworthy that the nature of the threats and opportunities presented by AI can be determined according to stakeholder groupings, with important distinctions across Faculty, Administrative Staff and Students identified (Hanover Research, 2023).

It must be emphasised that clear opportunities for the enhancement of teaching and learning have also been identified over the reporting period (Nerantzi et al 2023). Although not the subject of this case study, these applications have significance. Not least, in how AI is presented definitionally to encompass the full range of its application and in avoiding any reductive association with threat (UNESCO, 2023). The current sectoral approach can probably be best summarised as follows. While the integration of AI technology into third level education can be viewed positively and negatively; and while concerns persist on the reliability and ethical implications of AI; institutions have avoided outright bans and have instead opted "to offer guidelines and training to faculty and allow them to determine whether and how to integrate AI into their classrooms and assignments" (Hanover Research, 2023, p. 2).

While all these factors were considered, a more compartmentalised response was also needed. In this case that centred on the threat posed by Generative AI to academic integrity and the counteractive capacity of existing technology. One of the most significant findings in this context was that traditional plagiarism detection tools could be circumvented (Khalil and Er, 2023). Over the reporting period, research findings became more nuanced; indicating that as Generative AI software became more sophisticated (e.g. GPT 3.5 versus GPT 4) detection tools became less effective. It was also clear that the prevalence of false positives, which varied considerably according to provider, decreased the reliability of detection software as it struggled to identify human-written control responses from AI generated data (Elkhatat, Elsaid and Almeer, 2023).

Effectively, the reliance of the sector on plagiarism detection technologies left it exposed to generative technologies. Previously plagiarism detection was the exclusive remit of the assessor and reliant on their knowledge and judgement. The exponential expansion of, and access to, subject matter knowledge bases reduced the feasibility of this traditional approach while inflating the success of and reliance on the technological alternative. In turn, QAE processes have formalised this reliance on technology in maintaining academic integrity. Primarily, this was because the technology provided demonstrable, verifiable and incontrovertible evidence that plagiarism had occurred by tracking similarity to external source documents. During the College's assessment cycle following the launch of Chat GPT, it became evident that this was not transferable to AI Detection.

It became apparent that even when existing technology detected AI usage, it could not provide the same level of verifiability and was therefore potentially controvertible. The College's Moderation Committee identified two instances where the reliability of the AI detection technology was questionable, one relating to a suspected false positive and the other to a suspected false negative. Specifically, the question arose as to why one paragraph in an essay was detected as AI generated when it was stylistically and tonally indistinguishable from the remaining paragraphs that were not identified. In the second instance, content in an essay that had been flagged by the primary assessor and which the Moderation Committee suspected as AI generated was not detected by the software indicator. This was partly explained by an elaboration from the provider on how the detector software works.

> We strive to maximize the effectiveness of our detector while keeping our false positive rate - incorrectly identifying fully human-written text as AI-generated - under 1% for documents with over 20% of AI writing. In other words, we might flag a human-written document as AI-written for one out of every 100 fully-human written documents.
>
> To bolster our testing framework and diagnose statistical trends of false positives, in April 2023 we performed additional tests on 800,000 additional academic papers that were written before the release of ChatGPT to further validate our less than 1% false positive rate.
>
> In order to maintain this low rate of 1% for false positives, there is a chance that we might miss 15% of AI written text in a document. We're comfortable with that since we do not want to incorrectly highlight human-written text as AI-written. For example, if we identify that 50% of a document is likely written by an AI tool, it could contain as much as 65% AI writing. (Turnitin, 2024, n.p.)

The College's findings substantially aligned with the eventual publication of recommendation by the NAIN (2023, p. 17) which concluded "Detection systems cannot be relied upon to detect use of GenAI accurately or consistently".

However, detection software's capabilities are somewhat secondary. What was more significant in this case, was the extent to which human judgement was relinquished in the transition to the reliance on technology to detect plagiarism and to what extent had this become normative academic procedure. It also needed to be established whether this shift had become similarly invoked in policies and procedures. It was also noted that a reliance on human judgement in AI detection was reaffirmed by the College's software provider. Previously, "similarity reports" were available to students prior to final submission of an assignment without instructor involvement. This was not the case with the AI Indicator because visibility and download access would only be available to the "instructor", with obvious implications. "The AI writing detection indicator and report are not visible to students. However, with the PDF download feature, instructors can download and share the AI report with students." (Turnitin, 2024, n.p.).

Case Outcome

It was found that the relinquishment of human judgement in the detection of non-AI generated plagiarism was limited because similarity reportage still required assessor oversight and review. This was reflected in policies and procedures and could be readily extended to apply in the case of Generative AI. However, it was also found that the demonstrable component of existing plagiarism detection software (i.e. the ability to produce verifiable evidence of wrongdoing) was no longer incontrovertible and that the burden of judgement had shifted to the assessor. The feasibility of this burden of judgement was also challenged by the increased sophistication of Generative AI and the difficulty in distinguishing human and artificially generated outputs. While Faculty maintained some confidence that they could identify the difference based on their knowledge of their students' work, the removal of incontrovertibility remained a complication. The QAE officer advised that the AI Indicator did not have full comparability with the existing technology (Similarity Report), could not be referred to in the same way and would need to be distinguished in policy and procedure documentation.

References

Allen Cu, M. & Hochman, S. (2023, Jan 22). *Scores of Stanford students used ChatGPT on final exams, survey suggests*. https://stanforddaily.com/2023/01/22/scores-of-stanford-students-used-chatgpt-on-final-exams-survey-suggests/

Deng, Y., Zhao, N., & Huang, X. (2023, December). Early ChatGPT User Portrait through the Lens of Data. In *2023 IEEE International Conference on Big Data (BigData)* (pp. 4770-4775). IEEE.

Elkhatat, A.M., Elsaid, K. & Almeer, S. (2023) Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text. *Int J Educ Integr 19*, (17). https://doi.org/10.1007/s40979-023-00140-5

Feuerriegel, S., Hartmann, J., Janiesch, C. & . Zschech, P., (2024). Generative AI. *Bus Inf Syst Eng 66*, 111–126 https://doi.org/10.1007/s12599-023-00834-7

Goldman Sachs (2023). *Generative AI could raise global GDP by 7%*. https://www.goldmansachs.com/insights/pages/generative-aicould-raise-global-gdp-by-7-percent.html

Hanover Research, (2023). *Benefits, Challenges, and Sample Use Cases of Artificial Intelligence in Higher Education*. https://www.insidehighered.com/sites/default/files/2023-10/Benefits%2C%20Challenges%2C%20and%20Sample%20Use%20Cases%20of%20AI%20in%20Higher%20Education.pdf

NAIN, National Academic Integrity Network, (2023). Generative Artificial Intelligence: Guidelines for Educators. QQI. https://www.qqi.ie/sites/default/files/2023-09/NAIN%20Generative%20AI%20Guidelines%20for%20Educators%202023.pdf

Nerantzi, C., Abegglen, S., Karatsiori, M. and Martínez-Arboleda, A. (Eds.) (2023). *101 Creative ideas to use AI in education*. DOI: 10.21427/V9CX-1J69

Khalil, M. & Er. E. (2023). Will chatgpt get you caught? rethinking of plagiarism detection. arXiv preprint arXiv:2302.04335.

Turnitin, 2024. AI Writing Detection Capabilities: Frequently Asked Questions. https://www.turnitin.com/products/features/ai-writing-detection/

UNESCO, 2023). ChatGPT and Artificial Intelligence in higher education. UNESCO. https://www.iesalc.unesco.org/wp-content/uploads/2023/04/ChatGPT-and-Artificial-Intelligence-in-higher-education-Quick-Start-guide_EN_FINAL.pdf