Teaching Case

JITTC

Journal of Information Technology Teaching Cases
2023, Vol. 0(0) 1–9


Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/20438869231178844
journals.sagepub.com/home/jittc
Sage

Evaluating the ethics of machines assessing humans The case of AQA: An assessment organisation and exam board in England

Isabel Fischer

Abstract

The case focusses on Assessment and Qualification Alliance, one of England's largest exam boards, and its evaluation of whether Artificial Intelligence (AI) could be deployed, in principle, for marking high-stakes assessments. At a time when generative AI, such as ChatGPT, has gained popularity, the case offers insights into the challenges and risks of algorithmic decision making and algorithmic fairness, such as accuracy and explainability. The case allows students to explore the role of ethics when developing an AI-based tool in an area that they all know very well: Most students will have had to sit high-stakes assessments in the past and are still likely to be assessed on an ongoing basis as a current student. Students will thus be able to relate to the case and have their own stake(s) and opinion(s) about whether they would want to be assessed by AI. The case is also more broadly applicable in raising general awareness of the challenges and potential risks involved when using AI for decision making and it encourages students to consider the wider consequences for all stakeholders that are triggered by the use of digital technology.

Keywords

artificial intelligence, algorithmic decision-making, artificial intelligence ethics, education, assessments, automated essay scoring

Introduction

In February 2023, Dr Cesare Aloisi, Head of Research and Development (R&D) at AQA, one of England's largest assessment organisation and exam boards[1] (see Appendix A), was preparing for an upcoming presentation on the ethics of using artificial intelligence (AI) to mark students' essay-based exams. The central theme of his presentation was going to be how to transition ethically from the paradigm of *humans marking humans* in exam situations, to *humans **and** machines marking humans*. Aloisi wondered if, and how the challenges and risks of AI-based decision-making, such as fairness, accuracy and explainability could be overcome.

AIEd: promising far-reaching solutions

Interest in AI in education (AIEd) was not new. In fact, it dated back at least to the 1960s. Over the years, the possibilities that this technology presented were so enticing, that by the late 2010s and early 2020s, there was growing belief that AIEd could bring substantial benefit to education and that it had the potential to transform the educational landscape worldwide (Nguyen et al., 2022). In the UK, organisations such as JISC (an organisation that focussed on digital transformation of tertiary education) and NESTA (an innovation hub) all shared the belief that the use of AI could be of great benefit to education, if used correctly. They saw AIEd as having the potential to reduce teacher workload,[2] improve consistency in marking, provide wide-scale personalised learning, and ensure greater consistency in the quality of learning provided by schools and other educational institutions across the UK (JISC, 2022; Baker et al., 2019).

JISC (2022) maintained that in the education system the impact of AIEd could be 'transformational'. AI could both extend capacity, by automating certain functions, and increase capability, by augmenting others (see Appendix B). Thus, there could be instances of automated marking

University of Warwick, Coventry, UK

Corresponding author:
Isabel Fischer, Warwick Business School, University of Warwick, Coventry CV4 7AL, UK.
Email: Isabel.Fischer@wbs.ac.uk

(machines marking humans) and augmented marking (humans and machines jointly marking humans). This offered the opportunity to harness the automation-augmentation paradox, with automation and augmentation co-existing, rather than being a trade-off between the two (Raisch and Krakowski, 2020). JISC had developed a model of AI maturity (see Appendix C) illustrating the potential impact of AIEd at different levels of maturity. At the transformational level, JISC believed that AI would free educators from routine administrative tasks and allow them to focus on engaging learners, and allow learners to have a fully personalised learning experience.

What is AI, AIEd and AES

AI was not 'one single thing', there were many techniques and applications that together were commonly grouped as AI – for example, Deep Learning and Natural Language Processing (see, for example, Jaffri, 2022). Many AI tools in education, that is, AIEd tools, including AES (Automated Essay Scoring) as one type of AIEd tool, used a mixture of non-AI rule-based statistical features and deep-learning algorithms and databases (e.g. Pytorch, Hugging face framework, and Transformer such as LongFormer). Exhibit 1 below illustrates some of the mixture of statistical features and deep-learning algorithms in AES.[3]

There were three broad areas in which AI is being used in education: (1) System-facing AI, providing information for managers and administrators; (2) Learner-facing AI, interacting with learners on an adaptive basis, with the aim of personalising the learning for each learner; (3) Teacher-facing AI, seeking to reduce teacher workload by automating tasks such as marking and assessment, detection of plagiarism and provision of feedback, as well as those providing insights about learner progress and helping teachers to experiment with different methods of teaching based on the AI-generated insights (Baker et al., 2019).

As AQA focussed on setting and marking of exams, the area of AIEd which interested Aloisi most was the teacher-facing activity, and in particular the field of automated essay scoring (AES). In this field, because AI could not get tired or bored, AI promised to increase grading consistency. AES also had the potential to prevent the 'tick and flick' approach, where the level and detail of feedback that markers gave, became less and less as the number of papers marked increased (Lewis, 2013, p.189). Aloisi and his colleagues at AQA noted that AES should not be seen as a homogeneous construct: two of the aspects worth considering were low-stakes versus high-stakes assessments; and short-text responses, used for example for language tests, versus longer-text responses, which required demonstration of both linguistic and substantive knowledge (see Exhibit 2).

Grading writing quality

AES systems had been around for a long time, with the first being developed in 1966. Project Essay Grade (PEG) as the system was known, had been developed to enable the College Board, an organisation based in the US that developed and administered thousands of standardised tests,[4] to streamline and speed up its essay scoring process (Dikli, 2006). PEG sought to grade the quality of the writing by looking for characteristics that were predictive of writing quality, such as essay length, diction, fluency, grammar and sentence construction. An experiment conducted in 1999 to test the accuracy of PEG concluded that it performed at least as well as human markers, and that it was extremely efficient, being able to grade approximately six documents per second (Shermis et al., 1999). The authors of the report concluded: 'The initial applications of automated text graders will be to provide assistance in the summative evaluation of written work. However, the automated text grading has its greatest potential in providing students with formative feedback about areas of strength and weakness' (p.7).

By 2023 the field of automated essay scoring and formative writing feedback had exploded. Advances in natural language processing (NLP) meant that besides Large Language Models (LLMs), such as ChatGPT (GPT = Generative Pre-trained Transformer), there were now numerous AI tools that could be used for formative writing feedback. Among them were Grammarly, MI Write,[5] Feedback Fruits,[6] Turnitin and Quill. Grammarly claimed that every day 30 million people and 50 000 teams around the world made use of its products (Grammarly, n.d.). Quill was being used by around 123,000 teachers in 28,000 schools (Quill, n.d.). Turnitin, which started out as a tool to help students and teachers identify plagiarism, had evolved and was now also used to provide writing feedback. Turnitin's products were being used by more than 34 million learners in more than 15,000 school and tertiary institutions across the world (Turnintin, 2019). In the US, one state used PEG as its sole method for providing state summative writing assessments and the system was being used for formative writing assessments in 1000 schools and 3000 public libraries across the US. The digital learning company, Pearson, which had also been using automated scoring since the 1990s and owned Intelligent Essay Assessor (IEA), maintained that as early as 2010, IEA been used to score millions of essays written by learners in grades 4 to 12 and in tertiary education. Pearson believed that IEA could be used in high-stakes exams, to provide a second opinion and to provide formative evaluations (Pearson, 2010).

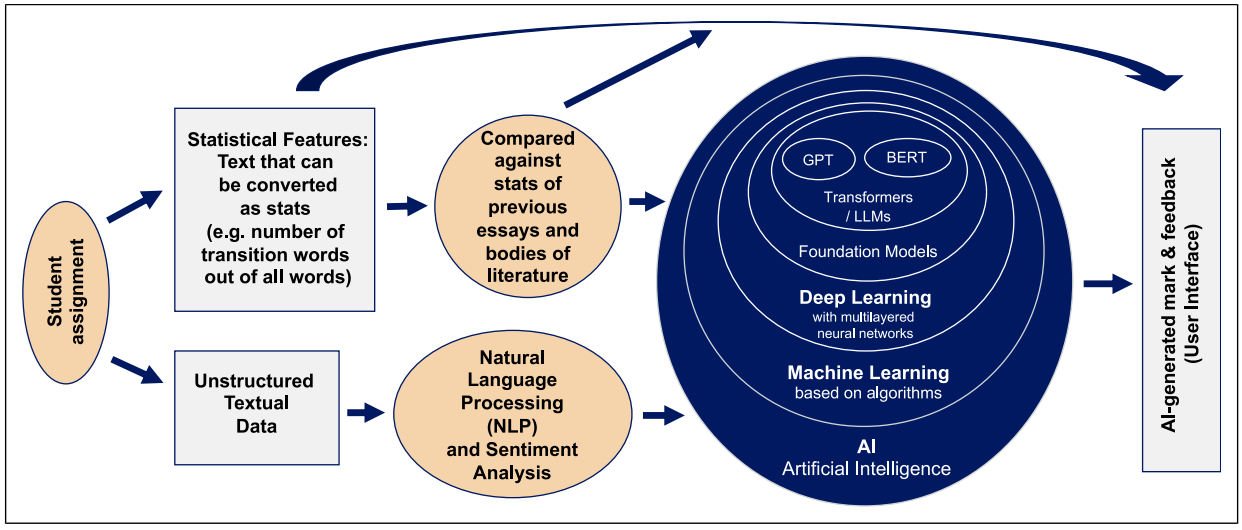

Exhibit 1. Statistical features and deep-learning algorithms in AES (simplified). Source: Adapted from Fischer et al. (2021).

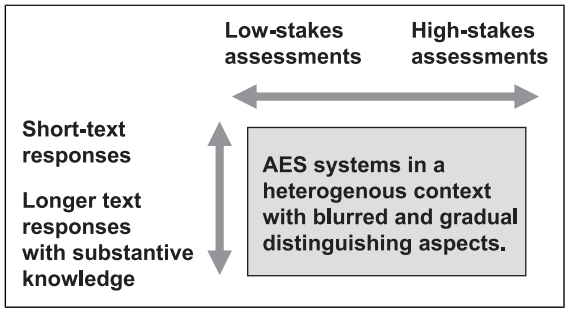

Exhibit 2. AES, a multi-dimensional construct. Source: Developed by Fischer and Aloisi for the purpose of this study.

The accuracy of AES

As Aloisi investigated AES, one of his key initial concerns was that of accuracy. Unlike human markers, AES systems did not evaluate the intrinsic qualities of an essay. Instead, they used 'correlations of the intrinsic qualities to predict the score of an essay' (Dikli, 2006). They used various techniques[7] to arrive at this prediction, but despite this, 'the basic procedure [was] the same. A relatively large set of pre-scored essays responding to one prompt [question] are used to develop or calibrate a scoring model for that prompt. Once calibrated, the model [was] applied as a scoring tool. Models [were] then typically validated by applying them to a second, but independent, set of pre-scored items' (Rudner et al., 2006). Depending on the AI applications used, some applications required as few as two pre-scored essays, others as many as 1000.

'In concept, a functioning model replicates the scores that would have been provided by all the human raters used in the calibration essay. Thus, a functioning model should be more accurate than the usual one or two human raters who typically assign scores', observed Rudner et al. (2006 p.18). 'The issue, however, is how one defines a validated functioning model... One never knows if the human or computer is more accurate. Nevertheless, one should expect the automated essay scoring models and humans raters to substantially agree and one should expect high correlations between machine and human-produced scores'.

Before approving the move to an AES product called Intellimetric, GMAC (Graduate Management Admission Council) had conducted research to assure itself that the tool would 'reasonably approximate' the scores of human markers. The evaluation had found the system to be 'extremely effective', and that it was even able to identify papers where cheating had occurred (Rudner, 2005). The agreement between Intellimetric and the human markers was very similar to that between two human markers – being identical or within one point of each other 97% of the time and identical 55% of the time (Kaplan, n.d.). The results of several other AES studies had also reported high agreement rates between AES systems and human assessors (e.g. Lewis, 2013 and Dikli, 2006), but of course, correlation between marks does not necessarily mean there is causation (Christodoulou, 2023).

Doubts and limitations on accuracy

Marjanovic and Cecez-Kemanovic (2017) and Galliers et al. (2017) pointed out that there were a number of limitations associated with algorithmic decision-making. These included the following:

- De-contextualisation: Data taken out of original context and then propagated and used in other contexts;
- Recombination: Creation of new data/information through re-combination of de-contextualised data from other sources;
- Using quantified proxies: Using quantified data as proxy measures for complex phenomena;
- Gaming: Strategic and selective collection and use of data in pursuit of individual goals
- Propagation of legitimation: Legitimacy of inferred information based on legitimacy of original data;
- Auditing by non-experts: Non-experts using open performance data judge the quality of complex expert activities;
- Amplified performativity: Data used to amplify impact of measures on what is being measured.

One of the above limitations that called into question the accuracy of AES was using quantified proxies. AES algorithms were frequently trained to identify words, phrases and patterns that were characteristic of stronger or weaker answers. They did not actually understand the essay that they were scoring. This raised the potential for users to mistrust the system and for the system to make mistakes. Indeed, at least two studies had shown that it was possible to trick certain AES systems by using a lot of big, but meaningless words (Lewis, 2013; Feathers, 2019). Other studies had found that even when as much as 20% of the content of an essay was changed, the AES score remained the same. On the other hand, simply adding three words to a 350-word essay could increase the AES score by an absolute 50% (Singla et al., 2021).

Christodoulou (2023) observed that once students know that AI is marking their essays, they want to know what it rewards and how it does so. They then try to game the system. She added: 'This, essentially, is the problem with AI marking. It's easy for it to be more consistent than humans, because humans are not great at being consistent, but whilst humans might not be consistent, they can't be fooled by tricks'.

For Aloisi, accuracy was especially important because of the grade boundaries in the high-stakes exams that AQA was involved in setting and marking. 'In a system like we have in the UK, where you have grade boundaries, one mark can make the difference between one grade and another grade', he said. He believed that it might be too much of a risk even to use an automarker as second marker in such high-stakes assessments.

Explainability

But accuracy was not his only concern in the early days of his research. One of his other main concerns related to the ethics of AES. 'Suppose that you have AI that is so good that it's indistinguishable from a human marker, what would we want to see to be able to trust it'? Aloisi asked. 'What emerged was explainability – can AI tell you why it gave a certain mark? The answer at the time was that it couldn't'.

This concern related to the 'black box' nature of AI, where the complex algorithms used in machine learning meant that the systems could arrive at conclusions that may agree with human conclusions, but were nevertheless unexplainable. Thus, in the case of AES, this made it difficult for humans to understand how these systems arrived at their conclusions. Even their creators sometimes found it hard to predict the conclusions that their systems would reach (Baker et al., 2019).

'Explainability is important for trust, because it gives you a sense that the system that you are interacting with is looking for more than superficial correlations: that it is capable of understanding some deeper meaning', said Aloisi. When thinking about trust, Aloisi used the ABI + model of trust, a model combining concepts of Ability, Benevolence, Integrity and Predictability to analyse the trustworthiness of a system (Aloisi, 2023). 'You want to know that the AI is aligned with you', he said. 'But also that it has the ability to evaluate you. If it's given you a mark, you want to know that the mark is based on some sort of academic judgement, as opposed to how many words you wrote or some other superficial thing. Trust is about making yourself vulnerable to someone else, because you think that person has your best interest at heart. If you have a system that cannot tell you why it gave you a certain score, to me, it's harder to claim that it's a trustworthy system'.

The reason why the human creators of AES systems could not explain how their systems arrived at their conclusions, noted Aloisi, was that 'the human will be able to tell you what the architecture is like, but they are not programming a set of rules. The system is designed in such a way that it can infer the rules. That is why it's called machine learning'. He likened the activity of trying to understand how AI came to its conclusions to the discipline of psychology, which seeks to understand why people behave in a certain way. 'Just looking at the brain and the way it is connected, you can have an idea of what's happening, because there are different areas of the brain that are associated with different things. But the processing side is still a massive area of learning', he explained. 'It's the same with machine learning – although much simpler. You know what the connections are, but you don't know for any given input, the sort of abstraction that it will make'.

Others had raised concerns arising from lack of explainability, one of which related to who could be held accountable for the conclusions reached by an AES. As one University College London professor put it: 'With humans there is accountability and exercise of power. What am I

going to do, fire the AI if it's incorrect? Who takes responsibility'? (Niemtus and Parker, 2022).

Another ethical concern related to the potential dehumanisation of learning. By their very nature, AIEd and AES systems sought to perform functions that were traditionally reserved for human beings. With education being so predicated on human interaction, some expressed concern at the consequences of removing humans from part of the process (Lewis, 2013; Comeau, 2019).

A third ethical concern related to the commercialisation and potential misuse of data. To date, most AIEd and AES systems had been developed by large corporations. Holmes (2022), for example, saw this as 'the commercialisation of education by stealth, as education systems increasingly rely on educational tools provided by the commercial sector'. For his part, Aloisi was not necessarily opposed to this commercialisation, but believed that there had to be a regulatory framework to facilitate this involvement.

A final ethical concern was that of bias and potential exacerbation of inequality. On the face of it, because AES did not involve a human marker, such systems had the potential to be completely unbiased. Lewis (2013) wrote: 'No human grader can be completely objective, even if the author of the essay is unknown. Certain writing styles and choices of topic or language can affect a human grader if only on a subconscious level. For a professor who interacts with students on a regular basis the possibility of bias entering into the grading process is a very real possibility. Favoured students are more likely to be graded leniently while out-of-favour students may be held to a stricter standard. A computer is not affected by such considerations'. However, there are still biases in AES. A study conducted in 2021 had shown a small, but significant, bias against male upper elementary school learners for AES. This bias was partly linked to essay word count. Removing word count did reduce bias marginally, but it also reduced each model's scoring performance (Litman et al., 2021).

ChatGPT and large language models a game changer?

Aloisi believed that the advent of large, pre-trained language models (also known as transformer-based models) was potentially game changing for AES. When he and his colleagues first started researching AES in 2018, these models did not exist, but in 2019, when large language models such as BERT (Bidirectional Encoder Representations from Transformers) started to appear, they started investigating the implications of these models for essay scoring. It was clear to them that these models were more accurate, and that they would continue to become more accurate as time went on. GPT (Generative Pre-trained Transformer)-3.5 and ChatGPT now showed potential to take this even further.

Christodoulou (2023) conducted an experiment using ChatGPT to test whether it was possible to game the system in the same way as it had been possible to game earlier AES systems. She found that while ChatGPT was wise to certain tactics, it did not pick up others, and she concluded that although it was hard to game the system, it was not impossible.

Aloisi remarked that these models seemed to address some of the issues of explainability. 'These days with generative AI, you can feed it an answer, you can feed it a mark scheme, and it will tell you a score and tell you why', noted Aloisi. He was not wholly convinced, however. 'It's not totally accurate. The explainability issue has not been completely resolved, but the systems have got better at giving explanations. In terms of how the system works, they are still black box systems. They are no more transparent than they were 5 years ago. They just crunch more data', he said.

'People can say that even people are black boxes: that they find it difficult to explain why they know something. But with people, you can keep probing. This is something that is only recently been made possible with ChatGPT. But what people have, that large, pre-trained language models don't have, is "direct experience of the world." We live in the world that we talk about. Whereas ChatGPT is only taught about the world. We can philosophise and say that even what we know about the world is mediated – I'm not claiming that humans are special in any way, it may just be a quantitative difference. But in my opinion, there is a huge quantitative gap between the way in which a person can access that academic judgement, compared to a piece of software'.

Large languages models could also be trained with very few papers, in what was known as 'few-shot', 'one-shot' and 'zero-shot' learning. 'That's why ChatGPT can work', observed Aloisi. 'You don't need that many essays to train the system'. But, he noted the accuracy of the systems diminished in zero-shot learning. 'Zero-shot learning means that you're getting accuracy of between 70% and 80%, which is fantastic from an R&D perspective, but if it's your children it's not good enough'.

He added: 'My argument is not that these things should not be used. My argument is that at the moment, because of the explainability issue, they are hard to scale in a high-stakes context, because you end up having to do so much quality assurance that you might as well pay a person to do it in the first place'.

'And I know that time will prove me wrong, because once you have the technology and people start to use the technology, *the exams will adapt to technology*. So the two things will start to co-exist, and it will become much easier. But at the moment, the adoption is starting from a stand-still, particularly in a high-stakes exam environment'.

EU AI Ethics Framework	EU Explanations	Broad areas (Jobin et al.)	Google
1 Human agency and oversight	Fundamental rights, human agency and human oversight	**Non-malefience**	Accountable to people
2 Technical robustness and safety	Resilience to attack and security, fall back plan and general safety, accuracy, reliability and reproducibility	*(underpinning varioius categories)*	Safe & Scientific excellence
3 Privacy and data governance	Respect for privacy, quality and integrity of data, and access to data	**Privacy**	Privacy design
4 Transparency	Traceability, explainability and communication	**Transparency**	*Transparency and explainability not explictly included*
5 Diversity, non-discrimination and fairness	Avoidance of unfair bias, accessibility and universal design, and stakeholder participation	**Justice and Fairness**	Avoid Unfairness / Biases
6 Societal and environmental wellbeing	Sustainability and environmental friendliness, social impact, society and democracy		Socially beneficial
7 Accountability	Auditability, minimisation and reporting of negative impact, trade-offs and redress	**Responsibility**	Accountable to people ctd. and 'Available for uses that accord with these principles'

Exhibit 3. Juxtaposing the EU AI ethics framework, academic literature and google. Source: Based on EU (2019), Jobin et al. (2019) and Google (n.d.)

Ethical guidelines

There seemed to be a global convergence around five principles for the ethical use of AI: transparency, justice and fairness, non-maleficence, responsibility and privacy (Jobin et al., 2019). Underpinning these principles seemed to be a consensus that complex normative questions could not be solved with 'good' design alone, and that while checklists made complex ethical debates appear straightforward, they did so in a conceptually shallow manner (Mittelstadt, 2019). Complexities included how imperfections in data might significantly impact AI-generated results and how the algorithms underpinning particular AI-based tools could be quite simple, however, the results too complex for the users (Rahwan et al., 2019).

In considering these issues, Aloisi believed that a possible starting point for developing robust AI-based assessments could be to identify the qualities of a good assessment and of a good assessor and to make sure that these qualities were inherent in AES systems. Furthermore, to get started, 'there are several AI ethics frameworks that can guide you', he pointed out. In the UK, a non-governmental Institute of Ethical AI in Education, established in 2018 to develop agreed principles for the ethical use of AIEd, had identified nine factors that should be taken into consideration when using AIEd (see Appendix D). Applicable to all AI, independent of industry sectors, the European Union (EU) had, in 2019, developed an AI Ethics Framework that was based on seven principles: human agency and oversight; technical robustness and safety; privacy and data governance; transparency, diversity and non-discrimination; and societal and environmental well-being. This framework aligns with broad areas identified by literature and also Google as an example form the private sector (see Exhibit 3).

The EU (2022) had taken the AI Ethics framework a step further and developed guidelines for the ethical use of **AI in education**. Four considerations were at the heart of the EU's AIEd guidelines:

1. **Human agency:** Ensuring autonomy, self-determination and responsibility;
2. **Fairness:** Treating all people fairly and ensuring that all have equal access to opportunity;
3. **Humanity:** Respecting the dignity, integrity and identity of all people and ensuring the 'well-being, safety, social cohesion, meaningful contact and respect that is necessary for human connection' (p?); and
4. Justified choice: Using 'knowledge, facts, and data to justify necessary or appropriate collective choices by multiple stakeholders in the school environment'.

> The EU said that this factor required 'transparency and is based on participatory and collaborative models of decision-making, as well as explainability'.

In addition, it had developed specific guidelines for the use of **AI for assessments (in education)**, recommending that the following factors should be considered before a school opted to use an AES system (EU, 2022):

- **Related to the concepts of diversity, non-discrimination and fairness:** Whether there are procedures in place to ensure that AI use will not lead to discrimination or unfair behaviour for all users;
- **Related to the principle of accountability:** Who will be responsible for the ongoing monitoring of results produced by the AI system and how the results are being used to enhance teaching, learning and assessment; and
- **Related to the principle of transparency:** Whether teachers and school leaders understand how specific assessment or personalisation algorithms work within the AI system.

Thinking it through

Aloisi reflected on research that said that the ambiguity and caution over the use of AIEd could be explained by the fact that it was at 'an emerging stage of hype, with over-optimism regarding the potential to transform existing education' (Humble and Mozelius, 2022 p.9). These authors had observed that a 90:10 phenomenon prevailed in AIEd, where 90% of the technology was working as it should, but the remaining 10% had the potential to cause the systems to fail. It seemed to him that this held true, but he had moved on from the perspective that these failings should completely prevent the use of AES in the kinds of settings in which AQA operated.

His thinking about AES had moved from explainability, bias, and reliability, to the question of what features and qualities were necessary for AI to work alongside people in an exam situation. 'I'm looking at trust, the components of trust, AI ethics, AI principles. The question for me is how we move incrementally from the paradigm of "humans assessing humans" to "humans with machines assessing humans"; how to make incremental changes that will make it easier to integrate AI technology into essay scoring; and how we can do this in an ethical way, so that people don't end up serving the machine'?

These were Aloisi's thoughts as he prepared his presentation for the forthcoming conference. He wanted to be able to make concrete recommendations and asked himself 'How can we do this'?

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Isabel Fischer https://orcid.org/0000-0001-7185-7579

Supplemental Material

Supplemental material for this article is available online.

Notes

1. Exam Board for A-levels and GCSEs (final, end of year exams, for UK secondary school students)
2. A 2019 Teacher Workload Survey conducted by Ofsted (the UK's Office for Standards of Education, Children's Services and Skills) found that there was 'more work to do to reduce unnecessary workload for teachers, middle leaders and school leaders' (Walker et al., 2019, p.14). A post-Covid teacher wellbeing survey conducted by the teachers union NASUWT, found that 90% of teachers had experience work-related stress in 2020 and that 52% said that workload had been the main reason for the increased stress (NASUWT, 2022).
3. In summary, AES (Automated Essay Scoring) is seen for the purpose of this case study as a subset of AIEd (AI in education), which in turn is a subset of AI.
4. Including the scholastic assessment tests (SATs) that many US students sit before being considered for entry into university.
5. See https://miwrite.com
6. See https://feedbackfruits.com/automated-feedback
7. For example, E-rater and Intellimetric used NLP techniques, Intelligent Essay Assessor used latent semantic analysis, No More Marking used comparative judgement.

References

Aloisi C (2023). The future of standardised assessment: validity and trust in algorithms for assessment and scoring. *European Journal of Education 58*(1): 98–110. https://onlinelibrary.wiley.com/doi/abs/10.1111/ejed.12542

Baker T, Smith L and Anissa N (2019), *Exploring the Future of Artificial Intelligence in Schools and Colleges*, NESTA https://media.nesta.org.uk/documents/Future_of_AI_and_education_v5_WEB.pdf (accessed 4 March 2023).

Christodoulou D (2023) *Can ChatGPT Mark Writing?* https://blog.nomoremarking.com/can-chatgpt-mark-writing-c98ff1f1a89 (accessed 4 March 2023).

Comeau R (2019), *TEACHER VOICE: Does AI, as a Tool, Deliver Student Feedback More Effectively than the Ballpoint Pen? the Possibilities – and Limits – of Artificial Intelligence*, The

Hechinger Report, https://hechingerreport.org/teacher-voice-does-ai-as-a-tool-deliver-student-feedback-more-effectively-than-the-ballpoint-pen/(accessed 4 March 2023).

Dikli S (2006). An overview of automated scoring of essays. *Journal of Technology, Learning and Assessment 5*(1). https://files.eric.ed.gov/fulltext/EJ843855.pdf

EU (2019) European commission. In: *Ethics Guidelines for Trustworthy AI*, Available at: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai (accessed 4 March 2023).

EU (2022) European commission, directorate-general for education, youth, sport and culture. In: *Ethical Guidelines on the Use of Artificial Intelligence (AI) and Data in Teaching and Learning for Educators*, Publications Office of the European Union, https://data.europa.eu/doi/10.2766/153756 (accessed 4 March 2023).

Feathers T. (2019), *Flawed Algorithms Are Grading Millions of Students' Essays*, Motherboard, https://www.vice.com/en/article/pa7dj9/flawed-algorithms-are-grading-millions-of-students-essays (accessed 4 March 2023).

Fischer I, Beswick C and Newell S (2021). Rho AI – leveraging artificial intelligence to address climate change: financing, implementation and ethics. *Journal of Information Technology Teaching Cases 11*(2): 110–116. DOI: 10.1177/2043886920961782

Galliers RD, Newell S and Shanks G (2017) Datification and its human, organizational and societal effects: the strategic opportunities and challenges of algorithmic decision-making. *The Journal of Strategic Information Systems 26*(3): 185–190. DOI: 10.1016/j.jsis.2017.08.002.

Google (n.d.) *Artificial Intelligence at Google: Our Principles*, https://ai.google/principles/(accessed 4 March 2023).

Grammarly (n.d.) *Our Mission*, https://www.grammarly.com/about (accessed 4 March 2023).

Holmes W (2022) *Edulearn22, Wayne Holmes Keynote Speech*, https://www.youtube.com/watch?v=fvhx-Cdd90I (accessed 4 March 2023).

Humble M and Mozelius P (2022) The threat, hype, and promise of artifcial intelligence in education. In: *Discover Artificial Intelligence*, file:///C:/Users/00100420/Downloads/s44163-022-00039-z%20(1).pdf (accessed 4 March 2023).

Jaffri A (2022) *Gartner Hype Cycle for Artificial Intelligence*, 2022. Stamford: Gartner, https://www.gartner.co.uk/en/articles/what-is-new-in-artificial-intelligence-from-the-2022-gartner-hype-cycle (accessed 4 March 2023).

Jisc (2022) *AI in Tertiary Education: A Summary of the Current State of Play*, https://repository.jisc.ac.uk/8783/1/ai-in-tertiary-education-report-june-2022.pdf (accessed 4 March 2023).

Jobin A, Ienca M and Vayena E (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence 1*(1): 389–399. DOI: 10.1038/s42256-019-0088-2.

Kaplan (n.d.) *What's Tested on the GMAT: Analytical and Writing Assessment*, https://www.kaptest.com/study/gmat/whats-tested-on-the-gmat-analytical-writing-assessment/(accessed 4 March 2023).

Lewis J (2013), *Ethical Implications of an Automated Essay Scoring (AES) System: A Case Study of Student and Instructor Use, Satisfaction and Perceptions of AES in a Business Law Course* https://digitalcommons.salve.edu/cgi/viewcontent.cgi?article=1047&context=fac_staff_pub (accessed 4 March 2023).

Litman D, Zhang H, Correnti R, et al. (2021) A fairness evaluation of automated methods for scoring text evidence usage in writing. In: Roll I (eds) AIED 2021, LNAI 12748, 255–267. DOI: 10.1007/978-3-030-78292-4_21 (accessed 4 March 2023).

Marjanovic O and Cecez-Kecmanovic D (2017) 'Exploring the tension between transparency and datification effects of open government IS through the lens of complex adaptive systems. *Journal of Strategic Information Systems 26*(3): 210–232.

Mittelstadt B (2019), Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence 1*: 501–507. DOI: 10.1038/s42256-019-0114-4.

NASUWT The Teachers Union (2022), *Teacher Wellbeing Survey*, https://www.nasuwt.org.uk/static/1ac040a7-96a5-481a-a052ddd850abc476/Teacher-Wellbeing-Survey-Report-2022.pdf (accessed 4 March 2023).

Niemtus Z and Parker K (2022) *Will a Machine Soon be Doing Your marking?* Tes Magazine, https://www.tes.com/magazine/teaching-learning/general/will-machine-soon-be-doing-your-marking (accessed 4 March 23).

Nguyen A, Ngo HN, Hong Y, et al. (2023) Ethical principles for artificial intelligence in education. *Education and Information Technologies 28*: 4221–4241. DOI: 10.1007/s10639-022-11316-w (accessed 4 March 23).

Pearson (2010) Intelligent essay assessor (IEA) fact sheet, http://images.pearsonassessments.com/images/assets/kt/download/IEA-FactSheet-20100401.pdf (accessed 4 March 23).

Quill (N.d.) No hype machine. Real data. Real impact. https://www.quill.org/impact (accessed 4 March 23).

Rahwan I, Cebrian M, Obradovich N, et al. (2019). Machine behaviour. *Nature 568*: 477–486.

Raisch S and Krakowski S (2020). 'Artificial intelligence and management: the automation-augmentation paradox'. *Academy of Management Review 46*(1): 1–48.

Rudner LM, Garcia V and Welch C (2005) *An Evaluation of IntelliMetric™ Essay Scoring System Using Responses to GMAT® AWA Prompts GMAC ® Research Reports* • RR-05-08 • October 26, gmac.com/~/media/Files/gmac/Research/research-report-series/RR0508_IntelliMetricAWA.pdf (accessed 4 March 23).

Rudner LM, Garcia V and Welch C (2006). An Evaluation of the IntelliMetricSM essay scoring system. *Journal of Technology, Learning, and Assessment*; *4*(4): Available from: http://www.jtla.org

Singla YK, Parekh S, Singh SS, et al. (2021) Automatic essay scoring systems are both overstable and oversensitive: explaining why and proposing defenses. Available from: https://arxiv.org/pdf/2109.11728.pdf (accessed 4 March 2023).

Shermis MD, Koch CM, Page EB, et al. (1999) Trait ratings for automated essay grading, https://files.eric.ed.gov/fulltext/ED432589.pdf (accessed 4 March 2023).

The Institute for Ethical AI in Education (2021), *The Ethical Framework for AI in Education*, University of Buckingham, https://www.buckingham.ac.uk/wp-content/uploads/2021/03/The-Institute-for-Ethical-AI-in-Education-The-Ethical-Framework-for-AI-in-Education.pdf (accessed 4 March 2023).

Turnintin (2019), Advance to acquire turnitin, leading provider of academic integrity, grading and feedback solutions, https://www.turnitin.com/products/features/draft-coach and https://www.turnitin.com/products/feedback-studio#what-can-you-do-with-feedback-studio-2 (accessed 4 March 2023).

Walker M, Worth J and Van Den Brande J (2019) *Teacher Workload Survey 2019 Research Report October*, National Foundation for Educational Research, https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/855933/teacher_workload_survey_2019_main_report_amended.pdf (accessed 4 March 2023).

Author biography

Isabel Fischer is Reader in Information Systems at Warwick Business School, UK, teaching in the area of digital innovation and AI ethics. Her research interest is at the intersection of education, technology and sustainability. Isabel joined academia after a career in digital payments.