November 2020
APPG AI Evidence Meeting

BIG INNOVATION CENTRE

AI in Education

Designing fair and robust AI-based assessments systems

PARLIAMENTARY BRIEF

AI in Education: designing fair and robust AI-based assessments systems *is a* Parliamentary Brief based upon the All-Party Parliamentary Group on Artificial Intelligence (APPG AI) Evidence Meeting held online on the 19th October 2020.

This Evidence Meeting was chaired by **Stephen Metcalfe MP** and **Lord Clement-Jones CBE**.

We would like to express our appreciation to the following people for their oral evidence:

- **Simon Buckingham Shum**, Professor of Learning Informatics, *University of Technology Sydney (UTS)*
- **Victoria Sinel**, AI Ambassador, *Teens in AI*
- **Cori Crider**, Co-Founder, *Foxglove*
- **Janice Gobert**, Professor of Learning Sciences & Educational Psychology, Graduate School of Education, *Rutgers University*; Co-Founder & CEO, *Apprendis*
- **Priya Lakhani O.B.E**., Founder CEO, *CENTURY TECH, Intelligent Learning, Artificial Intelligence in Education*
- **Laurence Moroney**, Lead AI Advocate, *Google*

Big Innovation Centre is the appointed Secretariat for APPG AI

- CEO, **Professor Birgitte Andersen**
- Rapporteur: **Dr Désirée Remmert**

The video recording of the Evidence Meeting can be found on our websites.

www.biginnovationcentre.com | Email: info@biginnovationcentre.com | @BigInnovCentre
www.appg-ai.org | Email: appg@biginnovationcentre.com | @APPG_AI

PARLIAMENTARY BRIEF

AI in Education: Designing fair and robust AI-based assessments systems

All Party Parliamentary Group on
Artificial Intelligence

APPG AI Sponsors

The Group supporters – Blue Prism, British Standards Institution, Capita, CMS Cameron McKenna Nabarro Olswang, Creative England, Deloitte, Dufrain, Megger Group Limited, Microsoft, Omni, Osborne Clarke, PwC, Rialto and Visa – enable us to raise the ambition of what we can achieve.

Contents

Introduction

Following the controversy around the flawed statistical model that was applied to standardise students' A-level results in England this summer, this APPG AI meeting took a closer look at different forms of AI-based assessment systems in education. Specifically, we explored the question of how these AI-technologies could contribute to a fairer assessment of students' performance at all levels of the education system. We discussed how these technologies can be deployed safely for students and teachers and how to ensure that they deliver accurate and transparent results. Further, we explored how AI-based assessment systems compare to teachers' assessment and how they may impact students' performance, motivation, and trust.

The APPG AI Evidence Meeting convened a group of experts in education research, technologists, teachers, and education policy.

- **Professor Simon Buckingham Shum**, Professor of Learning Informatics, *University of Technology Sydney (UTS)*
- **Victoria Sinel**, AI Ambassador, *Teens in AI*
- **Cori Crider**, Co-Founder, *Foxglove*
- **Janice Gobert**, Professor of Learning Sciences & Educational Psychology, Graduate School of Education, *Rutgers University*; Co-Founder & CEO, *Apprendis*
- **Priya Lakhani O.B.E.**, Founder CEO, *CENTURY TECH, Intelligent Learning, Artificial Intelligence in Education*
- **Laurence Moroney**, Lead AI Advocate, *Google*

This meeting was chaired by **Stephen Metcalfe MP** and **Lord Clement-Jones CBE.**

Parliament has appointed Big Innovation Centre as the **Secretariat of the APPG AI**, led by **Professor Birgitte Andersen (CEO)**. The Project Manager and Rapporteur for the APPG AI is **Dr Désirée Remmert**.

The evidence presented at the APPG AI meeting suggests that a **government-led initiative** is the **key measure** to ensure that the benefits of AI-driven assessment tools can be fully harnessed. This initiative **supported by educators, researchers, and developers,** fosters AI knowledge in **teachers, students, and the public. Stakeholders** must be **involved in the design and implementation** of these technologies to ensure their **applicability and user-friendliness**.

Further, the government must provide **adequate funding to schools and individual students** to make sure that there is **equal access to AI-tools in education across the entire education system.** Beyond this, the government must secure **regular independent auditing** of these technologies to guarantee their **fair and accurate deployment in classrooms.**

The brief will start with a discussion of how the AI-based assessment system can improve teaching at UK schools and universities. Here, we will look at some of the challenges that these technologies present - specifically, we want to explore how to guarantee that students' trust in a fair assessment of their performance will be preserved when applying AI-assisted assessment methods. We will then have a closer look at how AI-technologies might contribute to a restructuring of formal assessment in schools and at university. In particular, we will discuss the benefits of formative versus summative assessment and how an AI-assisted, less exam-oriented culture at schools and universities might positively affect students' performance and mental well-being. The brief will conclude with suggestions for policymakers on how to facilitate the fair and transparent deployment of AI-based assessment systems in education and summarise the benefits and challenges that the application of these technologies might mean to students, teachers, parents, and the wider society.

1. How can students and teachers profit from AI-based assessment systems?

The controversy around the **statistically adjusted A-level results** in England in August 2020[1] has given rise to a **broader public debate on the deployment of AI-technologies in education**. In this context, especially AI-assisted assessment technologies and their applicability at schools and universities in the UK have gained attention. However, it is critical to acknowledge that the statistical models that were used to generate the standardised grades are **fundamentally different from those specialised AI-assessment tools that will be discussed in this brief**. Nevertheless, the controversy might have amplified existing fears among students and teachers around AI-technologies in education that must be taken into account when discussing the deployment of AI tools at schools and universities.

Cori Crider, Co-Founder of Foxglove, states their concern about the way government adopted algorithmic decision-making in the context of the A-level exams in 2020. She criticises that the process of grade standardisation happened

> *"combined with a **lack of transparency, a lack of adequate public consultation, a***

[1] Zimmermann, Annette (August 14, 2020): "The A-level results injustice shows why algorithms are never neutral." *New Statesman*. Accessed November 5, 2020. https://www.newstatesman.com/politics/education/2020/08/level-results-injustice-shows-why-algorithms-are-never-neutral
Wakefield, Jane (August 20, 2020): "A-Levels: Ofqual's 'cheating algorithm' under review." BBC. Accessed November 5, 2020. https://www.bbc.com/news/technology-53836453

> ***lack of governance, and lack of understanding of the risks and limitations*** *inherent to the technology. This has led to serious and* ***costly failures*** *which have both* ***harmed the individuals*** *affected by the decisions and* ***wasted considerable sums of public money."***

Janice Gobert, Professor of Learning Sciences & Educational Psychology at Rutgers University, points out that the method with which the A-level results were adjusted was inherently flawed due to its reliance on aggregate data to predict the score of an individual student. Gobert explains:

> *"Specifically, an "AI algorithm" was developed that used an individual student's scores based on their teachers' grades in addition to an aggregate of each student's school data from prior years. However,* ***using aggregate data to predict the score of one student is problematic****, whether the method to do this was AI algorithm or traditional statistical techniques such as regression. Note* ***the issue here is NOT that AI was used to do this—rather, the problem is that aggregated data was used*** *to predict an individual's score(s)."*

Basing high stakes decisions on these data, she continues, increases the risk of errors as it will mostly **favour students whose scores are situated around the average** of the aggregate pool. The outliers, however, might see their grades either deflated or inflated.

Crider, drawing on an in-depth examination of the issue conducted by Foxglove, argues that "Ofqual's model was **irrational, arbitrary, failed to take account of relevant considerations**, and took into account irrelevant considerations." Specifically, she criticises that the model failed to consider teacher assessment in cohorts of more than 15, instead of relying on those assessments in cohorts under 5, which additionally skewed the results. Further, Crider asserts, the **entire process lacked transparency** which made it difficult to know how results had been generated:

> *"It was extremely difficult to access timely information about what the system was or how exactly it worked. On "results day", individual students were provided with* ***no clear explanation*** *of how their grades had been decided. This* ***undermined trust*** *and made it more difficult for students to understand or challenge decisions".*

Victoria Sinel, AI Ambassador at Teens in AI, delivered a moving account of how her A-level results had been negatively affected by an unfortunate interplay of Ofqual's standardisation method and a change of teachers in her class. **Receiving grades that were far below the average grades she had received before**, she had to **shelve her plans for further education** and repeat a year at college. Sinel reports:

> *"To say I was disappointed when I opened my results was an understatement, in fact to see that the A-Level algorithm had lowered every single one of my grades that my teachers gave me just based on the college I attended was demoralising. I'm now having to spend a third year in college because even one of the grades I got from a*

teacher was completely underestimated just because they didn't know me very well."

However, Sinel acknowledges that she is aware that the errors resulted from flawed models and a problematic methodology and argues that she would still trust tested AI-systems to deliver a reliable assessment of student performance in the future. Nevertheless, **the experience she and her peers have made with assessment technologies during their A-levels might leave a lasting negative impression that might be difficult to erase** as they move on into further and higher education.

Priya Lakhani OBE, Founder and CEO of Century Tech, expresses her concern that the controversy around Ofqual's exam standardisation methods has **undermined public trust and might bias students against data-driven technologies in assessment** in the future. Lakhani argues that Ofqual's model "despite not being a sophisticated use of AI, it still **failed to adhere to basic ethical principles**, such as impacting positively on learning, being fair and explainable to those it impacts, and using unbiased data." Lakhani fears that the **public mistrust** that the controversy caused might **set back data-driven technologies in education for years**. This would be especially unfortunate considering the many benefits that AI-driven assessment systems could offer teachers and students.

Simon Buckingham Shum, Professor of Learning Informatics at the University of Technology, Sydney, highlights some of the benefits that AI-based learning and assessment software could bring to education. He argues that **AI tutors, if used in an informed manner, can pinpoint students' strengths and weaknesses with great precision** and alert teachers to the areas where individual students will potentially need special attention. Predictive models can thus efficiently assist teachers in becoming "**more proactive, providing more timely support** to students before they drift too far off course." This will be especially **beneficial to larger classes** in which individual students usually do not receive sufficient personalised assistance from teachers.

Further, Buckingham Shum argues, **time-consuming types of assessment and feedback might be facilitated with AI-assessment technologies** which will result in more timely and detailed feedback for students and the alleviation of teachers' workload. For instance, learning analytics generated with activity data that is collected through students' interaction with learning technologies can provide students and teachers with **evidence-based predictions of their personal needs**. This way, they enable timely interventions that will contribute to a more effective learning experience for students (Lim et al. 2020: 2).[2] Recent research has even shown that personalised feedback from learning analytics software **elicits positive responses and can effectively increase motivation** in students (Lim et al. 2020: 16). The

[2] Lim, Lisa-Angelique Shane Dawson, Dragan Gašević, Srecko Joksimović , Abelardo Pardo, Anthea Fudge & Sheridan Gentili (2020): "Students' perceptions of, and emotional responses to, personalised learning analytics-based feedback: an exploratory study of four courses." *Assessment & Evaluation in Higher Education*, DOI: 10.1080/02602938.2020.1782831.

authors found that "[…] even though the feedback was automated to facilitate the provision of feedback at scale, students still recognised the role of the instructor in generating the feedback, as shown by the frequent comments pertaining to the perception of care" (ibid.).

Moreover, Gobert stresses, besides improved personalised feedback on their performance, novel forms of assessment that AI-driven systems can facilitate might also improve students' learning. She stresses that especially in science, **existing assessment methods would not prove ideal for assessing students' knowledge**. Gobert highlights the multiple problems that arise in the context of multiple choice and high-stakes summative tests.[3] While easy to grade, these methods do not offer any conclusion about the reasoning that underlies students' answers or are too general to provide an impression of a student's level of knowledge on a particular matter. Gobert explains: "Multiple-choice items are easy to grade, but these items are not authentic to the practices of science thus their ecological validity is very low, and they are not sufficient to assess the competencies outlined by international frameworks." Especially **summative tests** that are built around a multiple-choice structure, Gobert argues, would **produce many false positives** as students' have a 25% chance of getting the answer right.

However, despite all the benefits that AI-driven assessment and learning analytics can offer students and teachers, Buckingham Shum notes, it must be considered that **learning software cannot replace human warmth and connection.** Nevertheless, AI tools in education, especially those that take over time-consuming tasks like student assessment, can free up space in teachers' schedules to engage with students and foster the classroom community.

[3] See Janice Gobert's evidence statement in the appendix for a detailed list of assessment problems and literature references.

2. How could the deployment of AI-based assessment help us rethink exams in the UK?

AI-assessment systems are likely to bring important changes to the exam culture at schools and universities and might **initiate a fundamental transformation of the current assessment methods**. Lakhani emphasises that the benefits of AI technologies might bring a "beneficial cultural change to education" – if used properly. Specifically, she continues, the deployment of AI in student assessment could **instigate a move away from summative assessment**. Summative assessment comprises methods of performance evaluation in the end of a specific period. It sums up the taught material and assesses what the student has learned. Instead, AI could facilitate a **greater emphasis on formative assessment** which comprises regular, light-touch tests that convey an impression of a student's progress to teachers. Lakhani argues that

> *"[b]y using technology like AI, a shift towards formative assessment would reduce the exam season stress that blights the lives of teachers and students alike. It would provide a more accurate picture of the student's ability. It would reduce the distortive, destructive motives to 'teach to the test', helping children to learn more naturally and freeing up time and energy to focus on what children need to learn in order to truly thrive."*

Similarly, Buckingham Shum advocates for a **rethink of the current exam culture in the UK**. He argues that high stakes exams may become irrelevant as a method to convey teachers an impression of students' knowledge and abilities. The usefulness of exams to offer conclusions about students' performance would be particularly questionable due to the artificial conditions

under which the assessment is made. By contrast, **AI-driven learning tools powered by analytics, he notes, "can continuously assess students** as they are learning over extended periods, under diverse and more authentic conditions, providing a more robust picture of their ability." Buckingham Shum stresses that an automation of teaching, assessment and feedback would be **especially effective in STEM** (Science, Technology, Engineering, and Mathematics) subjects. He notes:

> *"There is great interest in **adaptive AI tutors that coach students at their own pace until they have mastered core skills and knowledge.** The full automation of teaching, assessment and feedback works for certain modes of teaching and learning, and where the student's mastery of the curriculum can be modelled in detail. In STEM subjects, there's evidence that compared to conventional learning experiences, students in school and university can learn more quickly and in some cases to a higher standard, using AI tutors."*

Nevertheless, Lakhani stresses, **AI-driven formative assessment and teacher oversight must go hand-in-hand**. Input and feedback are not mutually exclusive, she explains, and the insight of educators to students' abilities and progress must be part of AI-based assessment systems." Further, AI-based assessment is **not limited to quantitative data but can be supplemented with qualitative assessment by teachers**. Learning and assessment technologies operate with highly advanced methods, Gobert explains, which makes their assessment and predictions highly reliable if used correctly. For instance, **high fidelity log files that are generated through students' engagement in virtual inquiry**, can offer greater validity than multiple-choice tests. Thanks to methodological advances in computational techniques, Gobert stresses, AI-driven assessment software can provide further insight into students' learning processes. "Data-mined algorithms", she explains, "can handle the 4 v's in students' log data, namely **volume, veracity, velocity, and variability** […]. Lastly, data-mimed algorithms can be used to **assess students in real time, at scale, and to drive scaffolds to students while they learn, blending assessment and learning** so instructional time is not used for assessment." From this follows that the skilled application of AI-driven learning and assessment tools can present teachers with a more detailed and accurate picture of students' performance.

Predictive analytics might play a central role in the classroom of the future to help teachers take pre-emptive measures before individual students can fall behind. Predictive analytics **(PLA) "are a group of techniques used to make inferences about uncertain future events**. In the educational domain, one may be interested in predicting a measurement of learning (e.g., student academic success or skill acquisition), teaching (e.g., of value for administrations (e.g., predictions of retention or course registration)" (Brooks and Thompson

2017: 61).[4] Predictive models can support teachers with providing support to students in a timely manner**, taking proactive measures before students fall too much behind** can thus benefit the learning progress of individuals and the class. AI-driven assessment and feedback systems enable students to receive more detailed feedback than teachers can provide with the current teacher to student ration in schools. "**This pays off particularly in large classes**", Buckingham Shum explains, "and for student work that is time-consuming to grade and give good feedback on, examples of the latter being complex capabilities such as producing high quality academic writing, and face-to-face teamwork."

However, without a thorough introduction to these new technologies, there is a risk that teachers will not use the software to their full potential. **Educating teaching staff about these technologies is thus indispensable to the effective implementation** of AI tools in education. Further, research in this field warns that whereas these tools can provide rich insights into students' performance, they might also lead to an **"information overload"** that complicates the work of teachers and might have an adverse effect on their ability to assist students (Herodotou et al. 2019: 1274).[5] It is after all the **task of the teacher to transform insights delivered by learning technologies into interventions in the classroom** (ibid.). For this reason, the expert speakers at the APPG AI meeting agree, the introduction of AI-technologies into schools and universities must happen **in close collaboration with the teaching staff**. "We all know that we can be crushed or boosted by the way feedback is given to us. Designed and used well, AI can amplify all that we know about the provision of timely, actionable, personalised feedback, that both motivates and challenges" argues Buckingham Shum, "[…] students report a stronger sense of belonging when AI is used to expand the teacher's ability to give good personalised feedback to hundreds of students. But in a dysfunctional teaching culture, tools powered by analytics and AI are dangerous because of the speed and scale at which they operate."

[4] Brooks, Christopher and Craig Thompson (2017): "Predictive Modelling in Teaching and Learning." In Lang, Charles, George Siemens, Alyssa Wise, and Dragan Gašević (eds.) *The Handbook of Learning Analytics*, pp. 61-68. DOI: 10.18608/hla17.005.

[5] Herodotou, Christothea, Bart Rienties, Avinash Boroowa, Zdenek Zdrahal, and Martin Hlosta (2019): "A large-scale implementation of predictive learning analytics in higher education: the teachers' role and perspective." *Education Tech Research Dev* 67:1273–1306.

3. How to ensure the safe deployment of AI-based assessment systems in education: Suggestions for policymakers

A major step towards the development of safe and accurate AI-driven assessment tools for education is the **fostering of relevant R&D skills with targeted government-led initiatives**. Laurence Moroney, Lead AI Advocate at Google, advocates for the **adoption of training programmes at universities and schools to increase the pool of AI talent** in the population. Moroney notes that

> *"many of the recent technological leaps forward that created new economies succeeded because there was a critical mass of developers building applications using these technologies, driving innovation, launching new industries, and creating jobs, wealth and more. […] We've had AI before, but it failed, and the famous 'AI Winter' happened. One reason could be that while great papers were written, and great inventions were made, there was no real way for developers to build with this and get it into the hands of users. […] Ultimately, by having a critical mass of developers, we believe that AI will become as normalized in our lives as the Web or the Smartphone. That's ultimately what success will look like, and the key to this is broad, accessible, education."*

Likewise, Moroney stresses, **employers will need assistance in finding AI-talent**, especially those with expertise in machine learning, when it comes to the development of AI tools in education. He reports that

> *"often, employers want to use ML or AI in their products but have no idea about how to go about it. When hiring staff, there was a general sentiment that they didn't have internal skills to identify the correct people to help them, and often relied on anecdotal questions they found on the internet!"*

Offer foundational training for teaching staff to guarantee a safe deployment of AI assessment software in schools

However, not only is there an ever-increasing need for educational programmes to foster AI-developers and providing employers with assistance in identifying talent, but also **foundational training for teaching staff** is needed to guarantee a safe deployment of AI-based assessment software in schools and at universities. Buckingham Shum flags the **education of teachers as one of the major challenges to the safe and effective deployment of AI-driven software in the classroom**:

> *"When teachers are suitably trained, they use the tools well, and value the insights they gain to make better use of their time. However, upskilling teachers is all too often neglected and under-funded."*

Moreover, teacher training is indispensable to ensure that **the limits of AI technologies in education** are recognised. Teachers, as well as students, must be equipped to **over-rule an AI diagnosis**. They should be trained to **critically engage with the data generated by the software to know when human intervention is necessary**. Buckingham Shum stresses: "[…] reliable and fair outcomes depend on greater human agency […]. In fact, critiquing and teaching an AI tool is a powerful form of learning for students."

Involve educators in the shaping of AI tools to build trust in an AI enabled education system

Buckingham Shum also emphasises the importance of **involving educators in the shaping of AI tools to build trust in the system**. He thus backs the recommendation given by educators and education researchers in the APPG AI evidence meeting on the implementation of AI tools in the curriculum that was held in March 2020. Like our previous speakers, Buckingham Shum asserts that giving educators a voice will make them **feel respected and help them "becoming champions to their peers"**. In this context, Lakhani stresses the need for education of the benefits of AI-driven tools among both, the public and, more specifically, among educators, to **end a "technophobia" that risks having a detrimental effect on the quality of education in the future**:

> *"In addition to convincing the public as to the benefits that algorithms can offer, we must also embark in an education of the education sector to overcome the false fear-mongering idea that 'robots' will take over our schools. No other sector suffers from the same technophobia that persists in education. We must persuade the public that shying away from technology in education is effectively akin to sending our educators into the classroom without the tools they need to do their job."*

Engage the Public in open discussion on the use of AI technologies in education

Crider suggests encouraging **public engagement in AI technologies in education** through appropriate **Select Committees or APPGs (e.g. this APPG)** that can convene a series of **public events on "democratic engagement and government algorithms**". Here, members of the public can engage in an open discussion on "when and how algorithms used to 'assist' government policy are appropriate (and when not), how they must be designed to comply with equality law, and what safeguards are necessary". In order to guarantee that the software

deployed in class is up to current standards, Crider recommends,

> *"independent auditors (experts in statistical/algorithmic bias as well as educational assessment and attainment) (have) to audit the systems – at design stage and regularly after implementation. Auditors should scrutinise the extent to which people holding protected characteristics, such as nationality or race, are disadvantaged. (Disadvantage can arise through direct data or data which operates as a proxy for a protected characteristic.)"*

Further, she recommends that AI-based assessment systems must provide **"a reasoned, articulated decision" to individual students** who are "subject to any form of assessment, including factors that were weighed in the award of a grade or the decision to pass or fail".

Launch a nationwide scheme to provide all school children with electronic devices for AI supported and upgraded education

Moreover, Lakhani stresses, there is the urgent need for a **nationwide scheme that provides all schoolchildren with the electronic devices necessary to reap the benefits of an AI-supported education**. This investment, she stresses, would **pay for itself over time through the benefits for society accrued with the help of improved education**. "With no sign of COVID going away," Lakhani argues, "it seems sensible this is the right place to start to ensure learning continues, never mind assessment." Further, she notes, **future grading algorithms must be developed and applied with close adherence to ethical guidelines**. They must be transparent and developed under consultation with experts. For such grading system to be reliable, Lakhani points out, **more granular data on students' performance throughout their education must be provided** and, for this to be possible, the **general lack of technology infrastructure in all UK schools must be addressed**.

Summary

From this follows that a **government-led initiative**, supported by **educators, researchers, and developers**, which fosters AI knowledge in teachers, students, and the public is the key measure to ensure that the benefits of AI-driven assessment tools can be fully harnessed. **Stakeholders** must be **involved in the design and implementation** of these technologies to ensure their **applicability and user-friendliness**.

Further, the government must provide **adequate funding to schools and individual students** to make sure that there is **equal access to AI-tools in education across the entire education system.** Beyond this, the government must secure **regular independent auditing** of these technologies to guarantee their **fair and accurate deployment in classrooms.**

4. Evidence

Simon Buckingham Shum, Professor of Learning Informatics, University of Technology Sydney (UTS)

Thank you for the opportunity to contribute to the important work of this All-Party Parliamentary Group on Artificial Intelligence. I am Professor of Learning Informatics at the University of Technology Sydney, prior to which I was a professor at The Open University in the UK. Over the last decade, I have been active in shaping the emerging field of Learning Analytics, co-founded the Society for Learning Analytics Research, and have published extensively on the human-centred design of educational technology powered by analytics and AI, with specific attention to the skills and dispositions learners need for lifelong learning.

WHAT ARE THE BENEFITS AND CHALLENGES OF DIFFERENT TYPES OF AI-BASED ASSESSMENT SYSTEMS IN EDUCATION?

There is great interest in adaptive AI tutors that coach students at their own pace until they have mastered core skills and knowledge. The full automation of teaching, assessment and feedback works for certain modes of teaching and learning, and where the student's mastery of the curriculum can be modelled in detail. In STEM subjects, there's evidence that compared to conventional learning experiences, students in school and university can learn more quickly and in some cases to a higher standard, using AI tutors [1-4].

A different class of technology has no deep knowledge of the curriculum or students' expertise but can still predict if a student is going to struggle academically [5, 6]. Student-support teams skilled in the use of predictive models are improving outcomes for struggling university students by making more timely interventions [7-10].

I will flag two challenges for the way we introduce these tools.

A key factor is teacher training. When teachers are suitably trained, they use the tools well, and value the insights they gain to make better use of their time [3, 10]. However, upskilling teachers areare all too often neglected and under-funded.

Secondly, while AI tutors can enable impressive gains in the efficiency of learning core skills and facts — what do we do with the time this release in the curriculum? Do we fill those free slots with more disciplinary knowledge and skills to master? A smarter strategy is to enrich the curriculum with activities to more fully develop the qualities that so many educationalists and employers are calling for: curiosity, collaboration, reflection, critical thinking, ethical thinking, systems thinking, holding perspectives in tension, and the readiness to step out of your comfort zone [11-15]. The frontier challenge is to harness analytics and AI to build the knowledge, skills and dispositions needed for lifelong learning, and a workforce better prepared for change and complexity. It is more challenging for AI to help build these higher-order qualities, but progress is being made [15, 16].

HOW CAN IT BE GUARANTEED THAT AI ASSESSMENT WILL DELIVER RELIABLE AND FAIR RESULTS?

Designing valid, reliable assessments is an established discipline, and AI should be held to the same standards. Some AI tutors are validated assessment tools, predictive of student performance in established exams [17-19]. Looking to the future, however, high stakes exams may become irrelevant as a yardstick, since they test students for just a few hours under artificial conditions [20]. Learning tools powered by analytics and AI can continuously assess students as they are learning over extended periods, under diverse and more authentic conditions, providing a more robust picture of their ability.

In the many contexts where full automation of teaching and assessment is not possible, AI can still give formative feedback. However, reliable and fair outcomes depend on greater human agency: both teachers and students must be equipped to question and over-rule an AI diagnosis [9, 21]. In fact, critiquing and teaching an AI tool is a powerful form of learning for students [22].

Finally, we must listen to educators. We know that when we give them a real voice in shaping AI tools, this builds trust in the system [23, 24]. They feel respected as professionals, and become champions to their peers [25].

How might AI-based assessment systems change the teacher-student relationship?

The skilled use of AI tutors shows teachers with much greater precision of how their students are doing. They can focus attention on what is proving the most difficult material [3]. Predictive models can help teachers become more proactive, providing more timely support to students before they drift too far off course [9].

Students can now receive feedback that in certain contexts is more timely and detailed than any teacher can provide [26]. This pays off particularly in large classes [10, 27], and for student work that is time-consuming to grade and give good feedback on, examples of the latter being complex capabilities such as producing high-quality academic writing [28-30], and face-to-face teamwork [31]. Chatbots are becoming increasingly common, and some people prefer to disclose more to an AI advisor than to a human because it's perceived as less judgmental. Students from minority groups have preferred to receive support from a pedagogical agent, which they feel is less biased towards them than human staff [32].

AI also opens new possibilities for teacher professional development to improve how they interact with students. For instance, movement sensors can reflect back to teachers how they are moving around the classroom as they teach, to provoke reflection [33].

So, while the teacher/student relationship will change, it remains fundamental. No AI is going to provide the warmth and support a student needs when they arrive on a Monday morning after a tough weekend in a broken home. There remains plenty for teachers and students to work on that will remain invisible to the machine.

How will these technologies affect students' motivation and trust in a fair evaluation of their performance?

We all know that we can be crushed or boosted by the way feedback is given to us. Designed and used well, AI can amplify all that we know about the provision of timely, actionable, personalised feedback, that both motivates and challenges [27]. For instance, students report a stronger sense of belonging when AI is used to expand the teacher's ability to give good personalised feedback to hundreds of students [10]. But in dysfunctional teaching culture, tools powered by analytics and AI are dangerous because of the speed and scale at which they operate.

Concluding remarks

We are at a pivotal moment. There should be no sense of inevitability about the way that AI in education unfolds. It's not magic — it's conceived, funded and built by people — who as we

speak are making design decisions about products that our schools, universities and businesses will soon buy. We need strategy and investment to ensure that AI shapes education in the most productive directions. This begs the fundamental question: What kind of learners does society need, to tackle our most intractable challenges? We cannot meaningfully discuss the future of AI in education, without discussing what kind of education we want.

SOURCES

1) Lovett, M., Meyer, O. and Thille, C. (2008). The Open Learning Initiative: Measuring the Effectiveness of the OLI Statistics Course in Accelerating Student Learning. Journal of Interactive Media in Education, 14, 1-16. https://jime.open.ac.uk/articles/10.5334/2008-14/galley/352/download/
2) ASSISTments: Research Impact & Efficacy https://new.assistments.org/research
3) Murphy, R., Roschelle, J., Feng, M., et al. (2020). Investigating Efficacy, Moderators and Mediators for an Online Mathematics Homework Intervention. Journal of Research on Educational Effectiveness, 13, 2, 235-270. https://doi.org/10.1080/19345747.2019.1710885
4) Koedinger, K. R. and Aleven, V. (2016). An Interview Reflection on "Intelligent Tutoring Goes to School in the Big City". International Journal of Artificial Intelligence in Education, 26, 1, 13-24. https://doi.org/10.1007/s40593-015-0082-8
5) Brooks, C. and Thompson, C. (2017). Predictive Modelling in Teaching and Learning. The Handbook of Learning Analytics Research (Chap.5). Society for Learning Analytics Research. https://doi.org/10.18608/hla17.005
6) Aguiar, E., Lakkaraju, H., Bhanpuri, N., et al. (2015). Who, When, and Why: A Machine Learning Approach to Prioritizing Students at Risk of Not Graduating High School on Time. In Proc. 5th International Conference on Learning Analytics & Knowledge. https://doi.org/10.1145/2723576.2723619
7) Herodotou, C., Hlosta, M., Boroowa, A., et al. (2019). Empowering Online Teachers through Predictive Learning Analytics. British Journal of Educational Technology, 50, 6, 3064-3079. https://bera-journals.onlinelibrary.wiley.com/doi/abs/10.1111/bjet.12853
8) Georgia State University: Graduate Progression Success Advising Program. https://success.gsu.edu
9) Herodotou, C., Rienties, B., Boroowa, A., et al. (2019). A Large-Scale Implementation of Predictive Learning Analytics in Higher Education: The Teachers' Role and Perspective. Educational Technology Research Devevelopment, 67, 1273–1306. https://doi.org/10.1007/s11423-019-09685-0
10) Lim, L.-A., Dawson, S., Gašević, D., et al. (2020). Students' Perceptions of, and Emotional Responses to, Personalised Learning Analytics-Based Feedback: An Exploratory Study of Four Courses. Assessment & Evaluation in Higher Education, 1-21. https://doi.org/10.1080/02602938.2020.1782831
11) National Research Council (2011). Assessing 21st Century Skills: Summary of a

Workshop. Committee on the Assessment of 21st Century Skills. Board on Testing and Assessment, Division of Behavioral and Social Sciences and Education. https://www.nap.edu/read/13215/chapter/1

12) ATC21S (2012). Assessment and Teaching of 21st Century Skills. University of Melbourne/Cisco/Intel/Microsoft. http://www.atc21s.org
13) Gardner, H. (2009). Five Minds for the Future. Harvard Business Review Press.
14) Deakin Crick, R., Huang, S., Ahmed-Shafi, A., et al. (2015). Developing Resilient Agency in Learning: The Internal Structure of Learning Power. British Journal of Educational Studies, 63, 2, 121- 160. http://dx.doi.org/10.1080/00071005.2015.1006574
15) Buckingham Shum, S. and Deakin Crick, R. (2016). Learning Analytics for 21st Century Competencies. Journal of Learning Analytics, 3, 2, 6-21. https://doi.org/10.18608/jla.2016.32.2
16) Joksimovic, S., Siemens, G., Wang, Y. E., et al. (2020). Beyond Cognitive Ability: Enabling Assessment of 21st Century Skills through Learning Analytics (Editorial). Journal of Learning Analytics, 7, 1, 1-4. https://doi.org/10.18608/jla.2020.71.1
17) Feng, M., Heffernan, N. T. and Koedinger, K. R. (2006). Predicting State Test Scores Better with Intelligent Tutoring Systems: Developing Metrics to Measure Assistance Required. In Proc. International Conference on Intelligent Tutoring Systems. https://doi.org/10.1007/11774303_4
18) Ritter, S., Joshi, A., Fancsali, S., et al. (2013). Predicting sStandardised Test Scores from Cognitive Tutor Interactions. In Proc. 6th International Conference on Educational Data Mining. https://www.educationaldatamining.org/EDM2013/papers/rn_paper_25.pdf
19) Feng, M. and Roschelle, J. (2016). Predicting Students' sStandardised Test Scores Using Online Homework. In Proc. 3rd ACM Conference on Learning@Scale. https://doi.org/10.1145/2876034.2893417
20) Luckin, R. (2017). Towards Artificial Intelligence-Based Assessment Systems. Nature Human Behaviour, 1, 3, 0028. https://doi.org/10.1038/s41562-016-0028
21) Kitto, K., Buckingham Shum, S. and Gibson, A. (2018). Embracing Imperfection in Learning Analytics. In Proc. 8th International Conference on Learning Analytics and Knowledge. https://doi.org/10.1145/3170358.3170413
22) Kirsty, K., Mandy, L., Kate, D., et al. (2017). Designing for Student-Facing Learning Analytics. Australasian Journal of Educational Technology, 33, 5. https://ajet.org.au/index.php/AJET/article/view/3607
23) Buckingham Shum, S., Ferguson, R. and Martinez-Maldonado, R. (2019). Human-Centred Learning Analytics. Journal of Learning Analytics, 6, 2, 1-9. https://doi.org/10.18608/jla.2019.62.1
24) Biswas, G., Segedy, J. R. and Bunchongchit, K. (2016). From Design to Implementation to Practice a Learning by Teaching System: Betty's Brain. International Journal of Artificial Intelligence in Education, 26, 1, 350-364. https://doi.org/10.1007/s40593-015-0057-9
25) Shibani, A., Knight, S. and Buckingham Shum, S. (2020). Educator Perspectives on Learning Analytics in Classroom Practice. The Internet and Higher Education, 46. https://doi.org/10.1016/j.iheduc.2020.100730

26) Pardo, A., Bartimote, K., Buckingham Shum, S., et al. (2018). OnTask: Delivering Data-Informed, Personalized Learning Support Actions. Journal of Learning Analytics, 5, 3, 235-249. https://doi.org/10.18608/jla.2018.53.15
27) Huberth, M., Chen, P., Tritz, J., et al. (2015). Computer-Tailored Student Support in Introductory Physics. PLoS ONE, 10, 9. https://doi.org/10.1371/journal.pone.0137001
28) Roscoe, R. D., Allen, L. K. and McNamara, D. S. (2019). Contrasting Writing Practice Formats in a Writing Strategy Tutoring System. Journal of Educational Computing Research, 57, 3, 723-754. https://journals.sagepub.com/doi/abs/10.1177/0735633118763429
29) Fiacco, J., Cotos, E. and Rosé, C. (2019). Towards Enabling Feedback on Rhetorical Structure with Neural Sequence Models. In Proc. 9th International Conference on Learning Analytics & Knowledge. https://doi.org/10.1145/3303772.3303808
30) Knight, S., Shibani, A., Abel, S., et al. (2020). AcaWriter: A Learning Analytics Tool for Formative Feedback on Academic Writing. Journal of Writing Research, 12, 1, 141-186. https://doi.org/10.17239/jowr-2020.12.01.06
31) Echeverria, V., Martinez-Maldonado, R. and Buckingham Shum, S. (2019). Towards Collaboration Translucence: Giving Meaning to Multimodal Group Data. In Proc. CHI Conference on Human Factors in Computing Systems. https://doi.org/10.1145/3290605.3300269
32) Richards, D. and Dignum, V. (2019). Supporting and Challenging Learners through Pedagogical Agents: Addressing Ethical Issues through Designing for Values. British Journal of Educational Technology, 50, 6, 2885-2901. https://bera-journals.onlinelibrary.wiley.com/doi/abs/10.1111/bjet.12863
33) Martinez-Maldonado, R., Mangaroska, K., Schulte, J., et al. (2020). Teacher Tracking with Integrity: What Indoor Positioning Can Reveal About Instructional Proxemics. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol., 4, 1, Article 22, pp.1-27. https://doi.org/10.1145/3381017

Victoria Sinel, AI Ambassador, Teens in AI

My name is Victoria Sinel and I am incredibly grateful to have been given an opportunity to speak about the topic of technology in education. I will outline my view of education, my experience with the A-Level algorithm and my opinion on AI assessments within the education system.

The education system has remained the same for a very long time. Students are still only learning to pass an exam and education, as it currently stands, is not equipping school leavers with any real world skills. It is pretty obvious now that the future is powered by and shaped by tech, so schools should be teaching more about technology. I'm not suggesting everyone needs to learn to code, I myself am not a coding enthusiast but I do understand how tech works and I understand the ethics behind AI. I am also well aware of the ways technology can go very wrong and the impacts it can have on us as a society. The negative impacts of algorithms have been very well documented and some examples include racial biases within the Google photo algorithm and Amazon's recruiting engine which was biased against women. I learned about this not in school, but in hackathons that I attend during weekends, but this should be taught in school.

Education in the UK does not favour people from disadvantaged backgrounds and if COVID-19 did anything, it highlighted this. Just 7% of the British population are privately educated and it comes at no surprise they are the ones who benefit from algorithmic assessment based systems. Using technology in a system that is currently so unfair will reinforce the societal imbalance and reinforce the system that is already so dysfunctional and so unfair.

Unfortunately, I was a part of the class of 2020. I have always been academically bright but on A-Level results day I was left feeling like a complete failure. I attended a state school for sixth form because private school is just not something my family was able to afford. My

college had an average reputation for grades but had I known the way 2020 was going to turn out, I definitely would have seeked out a much higher performing college to attend, even if it meant me having to travel further.

To say I was disappointed when I opened my results was an understatement, in fact to see that the A-Level algorithm had lowered every single one of my grades that my teachers gave me just based on the college I attended was demoralising. I'm now having to spend a third year in college because even one of the grades I got from a teacher was completely underestimated just because they didn't know me very well.

Comparing this to my friend who was privileged enough to attend a private school, she seemed to luck out. The algorithm predicted her to get top grades not only based on her academic strength but also based on the school she attended since private schools are known to be very high achieving. She is now on a gap year before she attends Harvard university next year.

All this algorithm did was reinforce the structural inequalities in our society because the wealthier, privileged part of the population once again rose on top of the average person who cannot afford to attend private schools. I know I am not only speaking for myself when I say this has made me lose trust in absolutely everyone involved: the government that spoke of the "mutant algorithm", my teachers as well as all those who were involved in developing this algorithm.

Ofqual stated that their model had about 60% predictive accuracy on average across A-level subjects, meaning that they expect 40% of grades would have been different had exams been sat, after testing the model on 2019 data. We were not just guinea pigs in a rushed government experiment, this was our futures; our doors to university, apprenticeships or even just straight into the world of work and now many of us are having to take different routes to achieve our goals. I cannot begin to explain to anyone who did not experience it how it felt to have studied for 13 years just to be failed by a rushed statistical algorithm.

Compared to the other speakers my knowledge on tech is very minimal as I am just starting out but the first thing I learnt when I attended Teens in AI hackathons was how important human-centred design is when it comes to solving problems that affect real life people. There will never be a perfect algorithm but working with the people who will be affected by this and have the most knowledge on this is the best way to try and eliminate as much bias and discrimination as possible.

Tech, as we all know, has the potential to help us in our lives but it can also do the opposite and create further imbalances and exacerbate the already existing inequalities in our society. Without talking to students, teachers, tech specialists and ethicists any AI based assessment systems will not be as fair as they should be.

Cori Crider, Co-Founder, Foxglove

About Foxglove

Foxglove is an independent not-for-profit organisation based in the UK, which was founded in 2019. Our team includes lawyers, technology experts, and communications specialists. We challenge the misuse of digital technology by governments and big tech companies and stand up for a future where such technology is used to benefit everyone.

Overview of our position on public sector algorithms

We are concerned that at present there is considerable appetite in government for the adoption of algorithmic decision-making, combined with a lack of transparency, a lack of adequate public consultation, a lack of governance, and lack of understanding of the risks and limitations inherent to the technology. This has led to serious and costly failures which have both harmed the individuals affected by the decisions and wasted considerable sums of public money.

The A-level results algorithm fiasco illustrated these problems well – but was not unique. Just a few weeks earlier, a separate legal challenge which we were supporting, against the Home Office, led to the abandonment of a visa streaming algorithm.

We believe there are some fundamental issues which need addressing before algorithmic decision-making should be deployed anywhere in the public sector, including for educational assessment. A failure to address these issues first will inevitably lead to further problems including repeats of the kind of public outcry and legal challenges which forced the abandonment of the A-level algorithm.

OUR SPECIFIC CONCERNS REGARDING THE A-LEVEL ALGORITHM

Our detailed legal arguments regarding the unlawfulness of the A-level algorithm can be seen on our website: https://www.foxglove.org.uk/news/grading-algorithm-judicial-review-letter

WE CHALLENGED THE DEPLOYMENT OF THE ALGORITHM ON FIVE GROUNDS.

Ground 1: Ultra vires

We argued Ofqual exceeded its statutory powers by developing a grading system which did not give a "reliable indication of knowledge, skills and understanding" and which did not "promote public confidence in regulated qualifications and regulated assessment arrangements".

Ground 2: Irrationality

We argued that Ofqual's model was irrational, arbitrary, failed to take account of relevant considerations, and took into account irrelevant considerations. For example by failing to give any weight to teacher assessment in cohorts of more than 15, whilst relying on those assessments in cohorts under 5, the same pupil, taking Maths and Further Maths, could have her grade in Maths marked down whilst being awarded a higher grade in the "harder" of the two subjects.

Ground 3: Breach of the General Data Protection Regulation ("GDPR") and the Data Protection Act 2018 ("DPA")

We argued that the algorithm breached the duty of fairness and the duty of accuracy, and subjected students to automated decision-making without appropriate authorisation, without adequate safeguards, and without an adequate right of appeal. Additionally, it appeared that Ofqual had failed to conduct an adequate Data Protection Impact Assessment (DPIA).

Ground 4: unlawful discrimination

We argued that the algorithm would disproportionately impact students with protected characteristics under the Equality Act 2010.

Ground 5: breach of procedural requirements

We argued that Ofqual had failed to conduct an Equality Impact Assessment, as required by the Equality Act, and failed to conduct a proper consultation about crucial parts of the algorithm (because significant policy changes were made after the public consultation).

These specific legal arguments were not tested in the courts because the government

abandoned the policy before they could be tested. Several of them, however, were later echoed in the Shadow Attorney General's analysis of the legal flaws of the algorithm.

BEYOND THESE SPECIFIC CONCERNS, WE HAVE FOUR FUNDAMENTAL CRITICISMS OF HOW THIS ALGORITHM WAS INTRODUCED:

1. Lack of democratic legitimacy

If a system is to be developed which affects the life-chances of millions of people, it should have some form of democratic mandate. There had been a worrying lack of explanation from the government of what they were proposing to do, let alone a proper public debate. The consultation process engaged only a rarefied cohort, and significant decisions were made after the consultation process. The depth and breadth of public opposition once it was known that an across-the-board decision had been made to algorithmically downgrade to avoid grade inflation speaks – eloquently – to the lack of democratic assent.

2. Lack of transparency

It was extremely difficult to access timely information about what the system was or how exactly it worked. On "results day", individual students were provided with no clear explanation of how their grades had been decided. This undermined trust and made it more difficult for students to understand or challenge decisions.

3. Unfairness and bias

The algorithm entrenched educational inequality, by tying the grades of individual students in 2020 to the historical performance of their schools. This was patently unfair, and example of the way algorithms can easily systematise bias and disadvantage the already disadvantaged.

4. Unaccountable

Accompanying the lack of transparency and fairness was a lack of a robust means of appeal or route to redress.

We are not naive about the limitations of other forms of assessment. Exams can favour certain kinds of pupils, coursework can be cheated, teachers can bring their own biases and prejudices into their assessments – not to mention the wider context of educational equality and longstanding attainment gaps. However, at least with other forms of assessment there have been opportunities to study and mitigate their shortcomings, and there is a level of public understanding and public buy-in to their use. Without proper safeguards, algorithmic decision-making risks entrenching forms of unfairness present in other forms of assessment, alongside introducing new ones.

We would recommend the following actions as a bare minimum before any future adoption of algorithmic assessment in education.

BY THE DEPARTMENT FOR EDUCATION:

1. Consult stakeholders, including teachers, students, educational experts, technical specialists, employers, and universities about any future plans to automate assessment.

Consultations should be public, recorded, and meaningful.

2. Engage independent researchers into 'unconscious bias' and 'automation bias', so that the ways human decision-makers relate to recommendation systems are more fully considered.

3. Transparently describe what forms of assessment are under consideration or development, what datasets are to be used and how they will be prioritised in any new system.

4. Allow independent auditors (experts in statistical/algorithmic bias as well as educational assessment and attainment) to audit the systems – at design stage and regularly after implementation. Auditors should scrutinise the extent to which people holding protected characteristics, such as nationality or race, are disadvantaged. (Disadvantage can arise through direct data or data which operates as a proxy for a protected characteristic.)

5. Ensure that any new system provides a reasoned, articulated decision to every pupil or student subject to any form of assessment, including factors that were weighed in the award of a grade or the decision to pass or fail.

6. Provide for regular, external, and public reviews of any redesigned grading process and account for the process to Parliament.

BY PARLIAMENT:

1. Secure comprehensive documentation of the abandoned A-level system and the abandoned visa algorithm and learn from their failings. Place this documentation into the Parliamentary record to further Parliamentary education and debate.

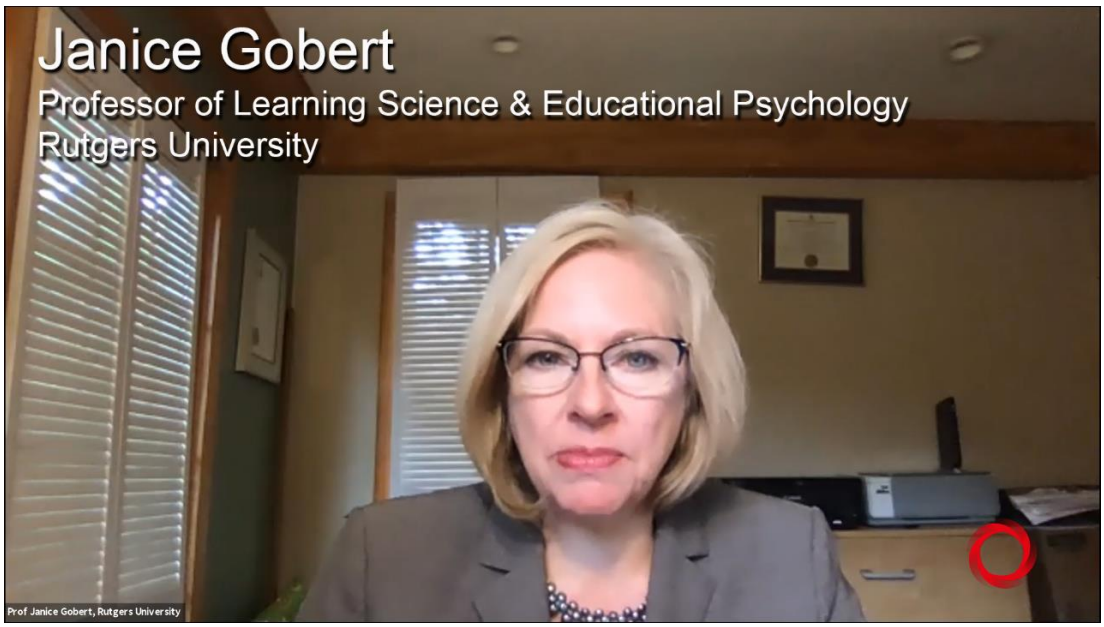
2. Through Parliamentary questions and other Parliamentary investigative techniques (e.g. Select Committee research), scrutinise the development of any new algorithmic decision-making systems in educational assessment.

3. Create a mandatory public register of algorithmic 'decision support' systems across UK public life and establish a formal Parliamentary process to scrutinise their use.

4. Encourage the government to implement the recommendations from The Committee on Standards in Public Life and Centre for Data Ethics and Innovation regarding the governance of algorithms.

5. Through appropriate Select Committees or APPGs (e.g. this APPG), convene a series of public events on “democratic engagement and government algorithms,” which ask the public when and how algorithms used to ‘assist’ government policy are appropriate (and when not), how they must be designed to comply with equality law, and what safeguards are necessary.

Janice Gobert, Professor of Learning Sciences & Educational Psychology, Graduate School of Education, Rutgers University; Co-Founder & CEO, Apprendis



I am a Professor of Learning Sciences and Educational Psychology at Rutgers Graduate School of Education (https://gse.rutgers.edu/janice_gobert) in New Jersey, USA.

I was trained as a Cognitive Scientist trained at the University of Toronto (Ph.D., 1994) and McGill University (M.A., 1989). My interdisciplinary research and development work sits at the intersection of Learning Sciences and Computer Science and focuses on computer-supported assessment and learning environments for the domain of science. My main focus for the past several years has been on the design and implementation of a science environment called Inq-ITS (inqits.com), and the assessment of students' processes at doing science, i.e., their competencies at conducting science investigations, as outlined in many policy documents (NGSS, 2013; UNESCO, 2019). As students conduct virtual inquiry with Inq-ITS' simulations, they are assessed and scaffolded in real-time using patented AI algorithms on all science practices (skills). Specifically, knowledge-engineering and educational data mining are used on students' logfiles (clickstream data) that are generated as they conduct inquiry with simulations within the system, and natural language processing algorithms score students' writing about their science investigations. (More details on Inq-ITS with respect to performance assessment is provided later).

Q1. How do assessment systems work and how they must be safeguarded to produce reliable results?

In the context of science learning and assessment (my area of expertise), there are many

problems with existing assessments typically used in school settings for both assessment and high stakes decisions, such as university placement. For example:

1) **Multiple choice / Fill-in blanks** measure factual knowledge but not the reasoning underlying students' thinking as they conduct science inquiry. Multiple-choice items are easy to grade but these items are not authentic to the practices of science, thus their ecological validity is very low, and they are not sufficient to assess the competencies outlined by international frameworks (e.g., NGSS, 2013; UNESCO, 2019).
2) **High-stakes summative tests** often rely on multiple-choice items that are too coarse-grained to accurately assess the breadth and depth of students' science knowledge and skills. Additionally, since students have a 25% chance of getting answers correct, there are many false positives with these items. Furthermore, high stakes tests like these cannot do true performance of students' competencies, and feedback is given too late to help students learn so students struggle in silence.
3) **Hands-on labs** are difficult to grade and keep consistent across tasks in terms of difficulty level, so measuring the development of students' skills over the school year is not easily accomplished.
4) **Lab reports** that require open response (possibly like the A level exams for science in the UK) are difficult to grade, and hand-scoring is subject to fatigue, bias, etc. (Myford & Wolfe, 2009). These also cannot fully capture science inquiry competencies, i.e., students' writing can lead to false negatives (students who can "do" inquiry experiments but cannot write about them), or false positives, (i.e. students who can parrot what they've read or heard but cannot successfully execute science experiments). Our research shows that between 30-60% of students are mis-assessed if teachers rely on students' writing for assessment purposes (Gobert, 2016; Gobert et al., 2019; Dickler et al., 2018, Li et al., 2018, 2019).

In addition to potential problems with various item types as outlined above, there is an additional problem with regard to how students' placement for university was done this year in the UK. Specifically, an "AI algorithm" was developed that used an individual student's scores based on their teachers' grades in addition to an aggregate of each student's school data from prior (3) years. However, using aggregate data to predict the score of one student is problematic, whether the method to do this was AI algorithm or traditional statistical techniques such as regression. Note the issue here is NOT that AI was used to do this—rather, the problem is that aggregated data was used to predict an individual's score(s). An added concern is that high stakes decisions were being made based on these aggregated data. Doing so has a high error rate. To wit, if a student's scores are typically higher than those of their peers, the score of the individual student is likely to be inaccurately deflated; whereas, if a student's scores are typically lower than those of their peers, the score of the individual student is inaccurately inflated. The only students who obtained accurate predictive scores using this method are those who were identical to the average of the students in the aggregate pool.

Q2. What are the benefits and challenges of different types of AI-based assessment systems in education?

I will situate my explanation of the benefits and challenges of AI-based assessment systems within the context of virtual science inquiry environments that do performance assessment of students' competencies at doing and writing about science, akin to what scientists do, and for which 21st century skills are needed for STEM careers (National Science Board, 2016).

The benefits are as follows. There are affordances of authenticity, and hence ecological validity when using an environment with simulations for inquiry because a simulation is perceptually very similar to hands-on apparatuses that would typically be used in class or by real scientists (Gobert, 2015). Specifically, with simulations, students can form a hypothesis, collect data, interpret data, warrant claims, and write an explanation of their findings (again, these skills or practices of science are reflected in national and international documents (NGSS, 2013; UNESCO, 2019), and are deemed as important to STEM careers.

As students engage in virtual inquiry, rich, high fidelity log files are generated, which can be used for performance assessment of students' competencies at conducting science. These offer greater validity than multiple-choice tests. Further, methodological advances in computational techniques, i.e., data mining offer analytical leverage on students' learning processes, not just products (Rupp et al., 2010). Data-mined algorithms can handle the 4 v's in students' log data, namely volume, veracity, velocity, and variability (Luan et al., 2020; Laney, 2001; Schroeck et al., 2012; Geczy, 2014). Lastly, data-mimed algorithms can be used to assess students in real time, at scale, and to drive scaffolds to students while they learn, blending assessment and learning so instructional time is not used for assessment.

Some considerations about the use of interactive environments for assessment purposes have been raised, though, in my opinion, these have been solved:

- Because complex tasks like science inquiry take longer, there can be fewer measures of "one type", and reduced reliability could result (Shavelson et al., 1999). Aggregating data across inquiry activities provides reliability (Mislevy et al., 2012; Mislevy et al., 2020).
- Because there is more than one way to conduct inquiry, there is variability in student responses. Note, educational data mining is particularly well suited to score the myriad of ways students conduct inquiry, both productively and unproductively (Gobert et al., 2013).
- In rich, multi-stepped tasks, the sub-tasks are not independent from each other, and thus, assumptions of conditional independence do not hold (i.e., Classical test theory).
- Traditional measurement methods tough to apply due to changing skill level as students learn in real-time (cf. Levy, 2012).
- Theory is needed to both distill/aggregate data (Quellmalz et al., 2012), and to design categories a priori so that results are pedagogically meaningful (Gobert et al., 2013). This can be solved by using both top-down and bottom-up means to

create coding categories, which in turn, inform the algorithms to be used (described next).

Q3. How can it be guaranteed that they will deliver reliable and fair results?

The design of activities is done both top-down and bottom-up design. Top-down design is done by using the literature on inquiry learning in science, and Evidence-Centered Design (Mislevy et al., 2012; 2020) to design tasks and technical tools to elicit students' inquiry competencies. Bottom up design is done via pilot testing both students and teachers, which helps to further define and reify the sub-components underlying inquiry (Gobert & Sao Pedro, 2017). As a result of using both methods, authentic science inquiry practices can be soperationalised and sconcretised for auto-scoring of science practices at scale (Mislevy, Yan, Gobert, & Sao Pedro, 2020).

Next, when implemented in real classrooms with diverse students, log data are collected, which is used to build and validate data-mined algorithms for assessment and scaffolding (Gobert et al., 2012, 2013). Briefly, text replay tagging of students' logfiles (Baker et al., 2006) is used to develop canonical models of what it means to demonstrate a particular skill, and in turn, develop and validate algorithms. Generalizability is tested with new students who were not used to build models (Paquette et al., 2014), ensuring they will deliver fair and reliable results for all students. It is also important to note that data-mining offers a considerable advantage regarding fairness in assessment over knowledge engineering approaches, which are solely derived top-down; these are less likely to be able to score the vast numbers of ways that students do inquiry, in part because these are subject to an expert blind-spot (i.e., it is difficult for experts to imagine all the ways in which students might conduct inquiry in a buggy manner).

Q4. How might AI-based assessment systems change the teacher-student relationship?

One recent innovation has great potential to change teacher-student interactions, namely, teacher alerting dashboards. Early findings are promising; for example, dashboards enable teachers to better monitor computer use in large classes (Chounta & Avouris, 2016), important in large classes and very useful during remote instruction due to COVID (Adair et al., submitted). Progress bars commonly used in dashboards allow teachers to quickly evaluate which students require immediate assistance and these can increase the number of interactions teachers have with students (cf., van Leeuwen, van Wermeskerken, et al., 2017). However, currently, many dashboards enable teachers to monitor only low-level indicators of student progress including the use of pedagogical resources and time spent on activities (e.g., Sosnovsky et al., 2012; Macfadyen & Dawson, 2010). Next I describe a dashboard based on AI algorithms, namely Inq-Blotter.

Inq-Blotter, for Teachers, paired with Inq-ITS, for students. Responding to teachers' and students' needs as well as policymakers' vision statements (cf., NGSS, 2013), our teacher

alerting tool, Inq-Blotter, provides alerts to teachers as to which students need help and on which inquiry practices and their respective sub-components. Inq-Blotter is fully integrated with Inq-ITS (Inquiry Intelligent Tutoring System; Gobert & Sao Pedro, 2017), in which students "show what they know" by conducting inquiry within microworlds/virtual labs (Gobert, 2015; see Figure 1). Unique to Inq-ITS is its capability to automatically assess students' inquiry using patented algorithms based on knowledge-engineering and data mining (Gobert, Baker, & Sao Pedro, 2014, US Patent nos. 9373082, 9564057, 10186168; Gobert et al., 2013, Mislevy et al., 2020). Inq-ITS was developed using Evidence Centered Design (Mislevy et al., 2012) so that the system elicits and collects evidence of students' proficiency at inquiry practices (Mislevy et al., 2012, Gobert & Sao Pedro, 2017). This allows the system to better capture what kinds of difficulties students have with inquiry, and then communicate those difficulties to the teacher with Inq-Blotter's succinct, formative, and actionable alerts (Gobert & Sao Pedro, 2017).

Inq-Blotter provides real-time technological support for assessment conversations (Duschl & Gitomer, 1997; Ruiz-Primo & Furtak, 2008) and closes the formative assessment loop and when can orchestrate learning (Prieto et al., 2011) both on a whole class level, when the teacher uses an alert to change instruction on the fly, or on an individual basis, when the teacher uses an alert to help individual students. This technology is transformative in terms of teachers' pedagogical practices, and in turn, students' inquiry learning. Specifically, our research has shown that what teachers say to students in responding to an Inq-Blotter alert can predict students' competency on the next inquiry task for the practice on which they were helped (Dickler, Gobert, & Sao Pedro, under revision).

Q5. How will these technologies affect students' motivation and trust in a fair evaluation of their performance?

There are many benefits to systems that offer real time support to students as they engage in high level tasks. In Inq-ITS, Rex, our digital agent supports students in real time when the system detects they need support. First, support is given in real time, when remediation of students' learning difficulties are most beneficial (Koedinger & Corbett, 2006); our system only jumps in when it detects that the student needs help (as opposed to on demand help, which can lead to help-abuse). Second, when a digital agent supports a student, his/her peers are not aware that a particular student was in need of help. Lastly, because students receive real-time help when they need it, they don't flounder too long, which can be frustrating and lead to disengagement (Gobert et al., 2015a, 2015b; Wixon et al., 2012; Rowe et al., 2009).

References

- Adair, A., Dickler, R., Gobert, J. & Lee, J. (submitted). Inq-ITS supports students maintaining their science inquiry competencies during remote learning due to COVID-19. Paper submitted for the 2021 Annual Meeting of the American Educational Research Association (AERA).
- Baker, R., Koedinger, K., Corbett, A.T., Wagner, A.Z., Evenson, E., Roll, I., Naim, M., Raspat, J., & Beck, J.E. (2006). Adapting to when students game an intelligent tutoring

system. In M. Ikeda, K. Ashlay, & T.-W. Chan (Ed.), Proceedings of the 8th International Conference on Intelligent Tutoring Systems, ITS 2006. LNCS 4053, pp. 392-401. Jhongli, Taiwan: Springer-Verlag.

- Chounta, I. A., & Avouris, N. (2016). Towards the real-time evaluation of collaborative activities: Integration of an automatic rater of collaboration quality in the classroom from the teacher's perspective. Education and Information Technologies, 21(4), 815-835.
- Dickler, R., Gobert, J., & Sao Pedro, M. (under revision). Using innovative methods to explore the potential of an alerting dashboard for science inquiry. Journal of Learning Analytics.
- Dickler, R., Li, H., & Gobert, J. (2018). False positives and false negatives in inquiry assessment: Investigating log and open response data. Presented at the European Association of Research on Learning and Instruction Sig 20 and Sig 26 2018 Meeting, The Hebrew University of Jerusalem, Jerusalem, Israel.
- Duschl, R. A., & Gitomer, D. H. (1997). Strategies and challenges to changing the focus of assessment and instruction in science classrooms. Educational Assessment, 4(1), 37-73.
- Geczy, P. (2014). Big data characteristics. The Macrotheme Review 3(6), 94–104.
- Gobert, J. (2015). Microworlds. In Gunstone, R. (Ed.) Encyclopedia of Science Education. Springer.
- Gobert, J. (2016). Inq-Blotter - A Real-Time Alerting Tool to Transform Teachers' Assessment of Science Inquiry Practices. Awarded by the National Science Foundation (NSF-IIS-1629045).
- Gobert, J., Sao Pedro, M., Betts, C., & Baker, R.S. (January, 2019). Inquiry skills tutoring system (child patent for additional claims to Inq-ITS). US Patent no. 10,186,168 (issued).
- Gobert, J., Sao Pedro, M., Betts, C., & Baker, R.S. (February, 2017). Inquiry skills tutoring system (child patent for alerting system). US Patent no. 9,564,057 (issued).
- Gobert, J.D., Baker, R.S., & Sao Pedro, M.A. (June, 2016). Inquiry skills tutoring system. US Patent no. 9,373,082 (issued).
- Gobert, J. D., Baker, R. S., & Wixon, M. B. (2015a). Operationalising and detecting disengagement within online science microworlds. Educational Psychologist, 50(1), 43-57.
- Gobert, J. D., Kim, Y. J., Sao Pedro, M. A., Kennedy, M., & Betts, C. G. (2015b). Using educational data mining to assess students' skills at designing and conducting experiments within a complex systems microworld. Thinking Skills and Creativity, 18, 81-90.
- Gobert, J. D., Sao Pedro, M. A., Baker, R. S., Toto, E., & Montalvo, O. (2012). Leveraging educational data mining for real-time performance assessment of scientific inquiry skills within microworlds. Journal of Educational Data Mining, 4(1), 111-143.
- Gobert, J. D., Sao Pedro, M., Raziuddin, J., & Baker, R. S. (2013). From log files to assessment metrics: Measuring students' science inquiry skills using educational data mining. Journal of the Learning Sciences, 22(4), 521-563.

- Gobert, J., Li, H., & Dickler, R. (2019). Dusting off the messy middle: Comparing students' science investigation competencies with their writing competencies. Presented at the 3rd International Conference on AI + Adaptive Education, Beijing, China.
- Gobert, J.D., & Sao Pedro, M.A. (2017). Digital assessment environments for scientific inquiry practices. In Rupp, A.A. & Leighton, J.P (Eds.), The Wiley Handbook of Cognition and Assessment: Frameworks, Methodologies, and Applications. West Sussex, UK (pp. 508-534).
- Koedinger, K., & Corbett, A. (2006). Cognitive Tutors: Technology Bringing Learning Sciences to the Classroom. In R. Sawyer, The Cambridge Handbook of the Learning Sciences (pp. 61-77). New York, NY: Cambridge University Press.
- Laney, D. (2001). 3D data management: Controlling data volume, velocity and variety. META Group Research Note, 6, 70-73.
- Li, H., Gobert, J., & Dickler, R. (2019). Assessing students' science inquiry practices and dusting off the messy middle. Poster paper presented at the European Science Education Research Association, Bologna, Italy.
- Li, H., Gobert, J., & Dickler, R. (2018). Unpacking why student writing does not match their science inquiry experimentation in Inq-ITS. In J. Kay & R. Luckin (Eds.), Rethinking learning in the digital age: Making the learning sciences count, 13th International Conference of the Learning Sciences (ICLS) 2018: (Vol. 3, pp. 1465–1466). London, UK: International Society of the Learning Sciences.
- Luan, H., Geczy, P., Lai, H., Gobert, J., Yang, S. J., Ogata, H., Baltes, J., Guerra, R., Li, P. & Tsai, C. C. (2020). Challenges and future directions of big data and artificial intelligence in education. Frontiers in Psychology, 11.
- Macfadyen, L. P., & Dawson, S. (2010). Mining LMS data to develop an "early warning system" for educators: A proof of concept. Computers & Education, 54(2), 588-599.
- Mislevy, R. J., Yan, D., Gobert, J., & Sao Pedro, M. (2020). Automated scoring in intelligent tutoring systems. In D. Yan, A. A. Rupp, & P. W. Foltz (Eds.), Handbook of Automated Scoring (pp. 403-422). Chapman and Hall/CRC.
- Mislevy, R.J., Behrens, J.T., DiCerbo, K.E., & Levy, R. (2012). Design and discovery in educational assessment: Evidence centered design, psychometrics, and data mining. Journal of Educational Data Mining, 4(1), 11-48.
- Myford, C. M., & Wolfe, E. W. (2009). Monitoring rater performance over time: A framework for detecting differential accuracy and differential scale category use. Journal of Educational Measurement, 46(4), 371-389.
- National Science Board (2016). Science and Engineering Indicators 2016 (NSB-2016-1), Arlington, VA: National Science Foundation.
- NGSS Lead States. (2013). Next Generation Science Standards: For States, By States. Washington, DC: The National Academies Press.
- Paquette, L., Baker, R. S., Sao Pedro, M. A., Gobert, J. D., Rossi, L., Nakama, A., & Kauffman-Rogoff, Z. (2014). Sensor-free affect detection for a simulation-based science inquiry learning environment. In Intelligent Tutoring Systems (pp. 1-10). Springer International Publishing.
- Prieto, L.P., Holenko Dlab, M., Gutiérrez, I., Abdulwahed, M., & Balid, W. (2011).

Orchestrating technology-enhanced learning: A literature review and a conceptual framework. International Journal of Technology Enhanced Learning, 3(6), 583-598.

- Quellmalz, E. S., Timms, M. J., Silberglitt, M. D., & Buckley, B. C. (2012). Science assessments for all: Integrating science simulations into balanced state science assessment systems. Journal of Research in Science Teaching, 49(3), 363-393.
- Rowe. J., McQuiggan, S., Robison, J., & Lester, J. (2009). Off-task behavior in narrative-centered learning environments. In, Proceedings of the 14th International Conference on AI in Education (pp. 99-106).
- Ruiz Primo, M. A., & Furtak, E. M. (2007). Exploring teachers' informal formative assessment practices and students' understanding in the context of scientific inquiry. Journal of Research in Science Teaching, 44(1), 57-84.
- Schroeck, M., Shockley, R., Smart, J., Romero-Morales, D., and Tufano, P. (2012). Analytics: The real-world use of big data. IBM Global Business Services.
- Shavelson, R., Wiley, E.W., & Ruiz-Primo, M. (1999). Note on sources of sampling variability in science performance assessments. Journal of Educational Measurement, 36(1), 61-71.
- Sosnovsky, S., Dietrich, M., Andrès, E., Goguadze, G., & Winterstein, S. (2012). Math-Bridge: adaptive platform for multilingual mathematics courses. 21st Century Learning for 21st Century Skills (pp. 495-500). Springer Berlin Heidelberg.
- UNESCO (2019). Beijing consensus on artificial intelligence and education. Unit for ICT in Education, https://enunesco.org/themes/ict-education
- van Leeuwen, A., van Wermeskerken, M., Erkens, G., & Rummel, N. (2017). Measuring teacher sensemaking strategies of learning analytics: A case study. Journal of Learning: Research and Practice, 3(1), 42-58.
- Wixon, M., d Baker, R. S., Gobert, J. D., Ocumpaugh, J., & Bachmann, M. (2012, July). WTF? detecting students who are conducting inquiry without thinking fastidiously. In International Conference on User Modeling, Adaptation, and sPersonalisation (pp. 286-296). Springer, Berlin, Heidelberg.

ACKNOWLEDGEMENTS

Inq-ITS and Inq-Blotter are generously funded by the US government's Dept of Education (R305A090170, R305A120778, EDIES15C0018, EDIES16C0014, 91990018C0022, 919900-19-C-0037) and the National Science Foundation (DRL-0733286, DRL-1008649, DGE-0742503, DRL-1252477, DRL-1643673, IIS-1629045)). All opinions expressed are those of the author and do not necessarily reflect the views of either granting agency.

Priya Lakhani O.B.E., Founder CEO, CENTURY TECH, Intelligent Learning, Artificial Intelligence in Education

1. For decades, artificial intelligence has been transforming every sector, from transport to healthcare. But in recent years, AI has emerged as an influential force within education. This has been primarily by way of personalising students' learning by rapidly and accurately learning their strengths, weaknesses and behaviours, and tailoring learning accordingly – ending the outdated 'one-size-fits-all' approach to education. Teacher facing tools have made use of AI and advanced data science to arm teachers with interventions data and administrators in educational organisations (particularly in higher education in the USA) have used AI technology to analyse academic analytics data within their institutions to drive efficiencies and predict dropout.

2. AI is now beginning to play an important role in other aspects of education – most notably, and recently most infamously, in assessment - although many experts in AI will make the point that given the algorithm was not 'learning autonomously', it was not in fact an AI - it was little more than a formula. The saga has arguably led to a public mistrust of algorithms. The use of an algorithm to decide students' A level and GCSE grades was an ill-thought out, counterproductive and harmful move that potentially set back the use of technology to improve education by years.

3. As co-founder of the Institute for Ethical Artificial Intelligence in Education, I noted that despite not being a sophisticated use of AI, it still failed to adhere to basic ethical principles, such as impacting positively on learning, being fair and explainable to those it impacts, and using unbiased data. In one particularly egregious example of the algorithm's harm, if no one from a student's school had achieved the highest grade in a subject in the previous three years, it was near-impossible for anyone from that school to be awarded that grade this year. This violates the principle of merit, which underlines our educational philosophy. This

algorithm placed more value on organisational data about the school (past performance) than it did place any weight on, for example, former assessment data of an individual student. Part of the challenge is that there is a lack of consistent granular data on a student's performance throughout their education. Often the sparse data in a school's management information system is based on qualitative data or summative assessment. The lack of technology infrastructure across all schools needs to be addressed. Future exam grading algorithms should closely adhere to ethical guidelines and should be developed with far more transparency, and in greater consultation with experts.

4. In a draft paper, Professor Rose Luckin of UCL says that in moving forward from this episode, the key to progress is "making sure that those designing the algorithm understand the context of application in appropriate detail and from a diverse set of perspectives. The right multi-stakeholder development group is needed for the design of any algorithm to be used for such a sensitive purpose, and that must include teachers, schools leaders, students." To this, I would add academics and AI experts and technologists.

5. But with all questions of policy making, we should be focused on the long-term. We shouldn't let the misuse of this algorithm prevent us from benefiting from its potential, when used correctly. The benefits of using AI in assessment properly could be system-wide, spurring on a wider beneficial cultural change in education. Using AI in assessment could facilitate a shift away from summative assessment – assessment that is used at the end of a period of learning to sum up what a pupil has learned, towards formative assessment – regular, light-touch tests that teachers give to understand what their pupils have learned. By using technology like AI, a shift towards formative assessment would reduce the exam season stress that blights the lives of teachers and students alike. It would provide a more accurate picture of the student's ability. It would reduce the distortive, destructive motives to 'teach to the test', helping children to learn more naturally and freeing up time and energy to focus on what children need to learn in order to truly thrive. Utilising AI as part of the end to-end educational process allows us to use technology to monitor progress in real-time, spot anomalies and intervene at the point of need, rather than every August when it is often too late. An AI-technology powered formative assessment system and teacher oversight, input and feedback are not mutually exclusive. Teachers' own insights into their pupils' abilities are invaluable and must be included in any assessment system. It is important to recognise that the data used for AI-based assessment systems does not have to be confined to quantitative data – teachers should be able to supplement data from summative assessment tests with their own qualitative assessments. This would not only improve the reliability and reduce the bias of results, but would ensure teachers are able to buy-in to the system. Combined, it offers a powerful solution to the sector.

6. When "one exam grade in every four is wrong", we must look at how we can use online assessment (incorporating data science or AI) to spot discrepancies earlier (including predictions of variances) or look towards a new form of assessment where an AI assessment system combined with learning technology can showcase a student's portfolio of work and achievements, challenging the need for high-stakes assessment.

7. AI could be involved in this proposed formative assessment system by using the data generated by a learning process that is personalised by AI – the data from learning feeds directly into the assessment system. For example, consider GMAT – the Graduate Management Admission Test used to assess entrants to business schools. It uses an algorithm that provides students with harder questions (and higher scores) when questions are answered correctly, and easier questions (and lower scores) when answered incorrectly. Through this relatively unsophisticated use of AI, the GMAT is able to assess a student's ability far more quickly than standard linear assessments. School assessment could be radically improved and streamlined using this approach – reducing workload and stress and allowing teachers to focus on educating their students.

8. In a more theoretical sense, AI could also be used to create a constantly-improving assessment system. As data on learning and assessment grows, data science can be used to improve the accuracy and reliability of algorithm-enabled grading year-by-year – a continually self-improving process.

9. But there are serious challenges to successfully using AI in assessment. Algorithmic bias (algorithms are programmed by humans and humans are inherently biased) – through which privilege to certain individuals or groups is hard-wired, producing unfair outcomes – plagues the use of AI in all sectors, from education to security to healthcare. This year, independent schools received double the increase in top grades as state schools, the result of the algorithm's bias towards smaller schools. Unequal access to technology would further harm the chances of the most disadvantaged – any online assessment system must be accessible to all students from all backgrounds. The pandemic has highlighted how the digital divide can harm educational outcomes. In the UK, almost one in ten families do not have a laptop, desktop or tablet at home – so relying on technology more for assessment could disproportionately harm those at the lowest rungs of the ladder of opportunity. We need to urgently discuss a nationwide scheme to provide all schoolchildren with a device. Is this not achievable in 2020? It is arguable this investment would pay for itself over time through the increased benefits to learning outcomes and its knock on effects to society (employment etc). With no sign of COVID going away, it seems sensible this is the right place to start to ensure learning continues, nevermind assessment.

10. On top of this, we must create a reliable, fair system that rewards effort as well as competence. The recent algorithm ignored the hard work of many students. In the Learning How to Learn APPG report there is a list of the soft skills society requires of the students of 2020. An AI assessment system allows us the opportunity to embrace and discover these skills within students. Such a system could also offer insights to ministries of education on where resources are required or further continuing professional development for teachers. An online real-time formative assessment system allows for instant action.

11. AI will never have a role to play in assessment until it can be trusted. Optimistically, some studies have shown that people adhere to advice more "when they think it comes from an algorithm than from a person" – contrary to the common belief that people instinctively distrust algorithms. Recent research suggests that people may trust AI systems over humans when

they are focused on utilitarian and functional outcomes, as opposed to experiential and sensory. Yet more research is required to determine the level of public trust in the use of algorithms in education. We also need to look at the practical applications of online assessment and combine other technologies (such as recognition software) to ensure such systems are not gamed by students, however, I would argue that these problems are not insurmountable and the technology exist to solve these issues.

12. Before adopting an AI-based assessment system, some more fundamental questions should be asked. What is assessment for? What do we count as success? Simply changing assessment to an online, AI-enabled system may not suffice. We have an opportunity to look at adopting a wider yet less bloated curriculum, encompassing the new skills, such as learning agility (see APPG report on Learning how to learn), that will be required to thrive in the age of automation and AI. And should we be endorsing a standardised system that effectively guarantees that a third of students will 'fail'?

13. The teacher-student relationship, focussed on enabling students to fulfil their potential, more one-to-one interaction and tailored support can be augmented – not replaced – by AI. A teaching, learning and assessment system that uses AI would save teachers vast amounts of time that could instead be used on teaching. Almost one in five teachers spends over 11 hours a week marking work; two thirds of teachers say their marking workload harms their ability to teach their students. If we are to treat education as a complex yet holistic system, then assessment must be approached with teacher workload – possibly the biggest crisis in education – in mind.

14. In addition to convincing the public as to the benefits that algorithms can offer, we must also embark in an education of the education sector to overcome the false fear-mongering idea that 'robots' will take over our schools. No other sector suffers from the same technophobia

that persists in education. We must persuade the public that shying away from technology in education is effectively akin to sending our educators into the classroom without the tools they need to do their job.

SIGNPOSTS FOR EVIDENCE:

Paragraph 6

https://rethinkingassessment.com/rethinking-blogs/just-how-reliable-are-exam-grades/

Paragraph 9

https://www.childrenscommissioner.gov.uk/2020/08/18/children-without-internet-access-during lockdown/

Paragraph 11

https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2941774

https://hbr.org/2020/10/when-do-we-trust-ais-recommendations-more-than-peoples
Paragraph 13

https://www.tes.com/news/workload-tens-thousands-teachers-spend-more-11-hours-marking
every-week

Laurence Moroney, Lead AI Advocate, Google

So many of the recent technological leaps forward that created new economies succeeded because there was a critical mass of developers building applications using these technologies, driving innovation, launching new industries, and creating jobs, wealth and more. In my career, I have seen two major ones: Starting in the 1990s, with the advent of the Web, the 30-year old internet finally had a platform that allowed developers to build applications to reach consumers. This led to the birth of the modern tech giants. Then, in the mid-2000s the launch of the smartphone -- a handheld device, that was internet-connected, offering a computing platform on which developers could put apps led to an explosion of innovation from the one-person start-up to the massive conglomerate. Again, whole new industries were born, and new wealth generated.

But without developers building innovative solutions around these platforms, none of these revolutions would have happened. The web needed Amazon, Google, MySpace and other applications to attract users to it. When the users came, others built for it, and the millions of web sites we know today became not just possible, but essential. The smartphone needed Uber, Instagram, WhatsApp and other applications to get users off their feature phones, and when they flocked to this platform in the billions, then other app ecosystems grew from it.

The key: developers. So, for AI to be a success, our vision is to empower developers everywhere to be successful with AI. These developers will build the apps that none of us have thought of yet, but which will create a whole new industry, and this one could be far larger than the Web or the Smartphone.

We've had AI before, but it failed, and the famous 'AI Winter' happened. One reason could be that while great papers were written, and great inventions were made, there was no real way

for developers to build with this and get it into the hands of users. I recall consulting for a startup in London in 1993 that was attempting to use Neural Networks to predict the usage of natural gas in the UK, that they could sell to British Gas so that they'd know how much to pump. It required custom hand-coding by a PhD for many months. Of course it failed.

In 2018, we read a paper that claimed there were 300,000 AI practitioners in the world. By most measures there are about 30 million software developers in the world. If we could teach only 10% of software developers how to build AI solutions effectively, we would increase the number of AI practitioners by 10x

SO OUR STRATEGY FOR DOING THIS BECAME:

The first thing is to produce a platform on which developers can build ML and AI solutions, without needing a PhD to understand it. That's what we call TensorFlow. It's free and open source, so there's no blocker there. It's available for anyone to use.

The second is to produce an education syllabus, and strategy to teach it developer-to-developer spanning MOOCs, Universities and other educational establishments. We launched this in 2019, and via MOOCs have trained about 600,000 people to date. But, in order for the skills to really take hold on a grassroots level, it should be people other than Google teaching it. A key to that is working with Universities, so in 2019, we also launched a plan to share our syllabus and materials with any university that wants it, and to help them, with funding, to get over the initial bumps in launching a new course. Some might need funds to pay for TAs. Some might need equipment or space. As such we launched a program that gave Universities that would teach TensorFlow a no-strings-attached financial gift. There was an application process, and institutions with the best applications, specifying what they would teach and how they would use the funds, were selected. In the UK we worked with Cardiff University, who launched a Masters in AI, with University College Oxford who are developing new courses, and Imperial College in London who created several courses, including an online MOOC.

The third part of our strategy was to help employers understand how to find people with applicable ML skills. We have identified that often employers want to use ML or AI in their products, but have no idea about how to go about it. When hiring staff, there was a general sentiment that they didn't have internal skills to identify the correct people to help them, and often relied on anecdotal questions they found on the internet! With that in mind, we launched a Developer Certificate exam, and associated social network, so that software developers could pass our rigorous tests, and then be able to show employers they have skills in the most sought-after areas of AI: Computer Vision, Natural Language Processing and Sequence Modelling. The goal is to break the logjam, and have the right people at companies, helping them with innovative solutions in AI. Additionally, we're in the process of launching a services partner network, so companies that want to take the next step -- and really go deep into AI and ML have a network of established, credible, services partners they can work with.

We're about 18 months into the plan, and we've already seen terrific results. We estimate that we've reached around 1 million developers and educated them in the basics of ML and AI, and from there the next wave of new applications, the ones that will use AI-as a platform the way Amazon used Web or Uber used smartphones and created entirely new industries, will grow from this. Recently we've seen programs such as one in Korea where a network of startups, needing AI skills, have launched a boot camp where, if one is selected, and one passes our courses, then they have a guaranteed job or internship

Ultimately, by having a critical mass of developers, we believe that AI will become as snormalised in our lives as the Web or the Smartphone. That's ultimately what success will look like, and the key to this is broad, accessible, education.

Contact

APPG AI Secretariat

Big Innovation Centre
62 Wilson Street
London EC2A 2BU
United Kingdom

info@biginnovationcentre.com
www.biginnovationcentre.com

appg@biginnovationcentre.com
www.appg-ai.org