The Institute for Ethical AI in Education

Developing a Shared Vision of Ethical AI in Education:

An Invitation to Participate

Contents

1. Introduction

The Institute's objective is to develop an ethical framework that will enable learners to benefit from artificial intelligence whilst being protected against its risks. If this framework is to have legitimacy and gain traction then it needs to be imbued with the views and values of those who stand to be affected by this technology. That is to say, it must represent a shared vision of ethical AI in education.

This report aims to facilitate this shared vision. In particular, it aims to inform and focus an open conversation, centred around a series of roundtable events that will take place throughout Autumn 2020. The Institute's final report, due to be published in March 2021, will reflect the discussions and conclusions from these roundtables.

To achieve its aim, this report presents:

- **A number of critical questions on what constitutes ethical practice**
- **A number of suggested methods for facilitating ethical practice, on which feedback is keenly invited**
- **A set of insights gained from a series of expert interviews, supplemented by secondary research to support participants' judgements**

If you are interested in AI in education, have relevant insights, or think that you will be affected in any way by the use of AI in education, please join the conversation. We would like to hear from learners of all ages, educators, academics, organisations developing and deploying AI, policymakers, influencers and concerned members of society. Please contact the Institute (see Section 7. Next Steps and Further Information) to tell us your:

- **Views about the critical questions summarised in Section 5**
- **Feedback on the Institute's suggested methods for facilitating ethical practice outlined in Section 6**
- **Feedback on further ethical issues and/or critical questions that should be considered**
- **Insights that may support others' judgements on the critical questions or other ethical issues**

2. Using Artificial Intelligence to Benefit Learners

AI has the potential to enhance teachers' practice and allow a greater proportion of their time to be reallocated to more effective tasks, such as coaching and mentoring.[1] AI has the propensity to provide educators with rich insights into the learning process, which may include greater levels of information on harder-to-measure skills and metacognition.[2] This could enable teachers to understand and meet the needs of learners better, and could also facilitate a shift away from high stakes assessments and towards more continuous assessment of learning. Moreover, AI has considerable potential to support learning directly. As described in a 2020 OECD working paper, "AI applications can identify pedagogical materials and approaches adapted to the level of individual students, and make predictions, recommendations and decisions about the next steps of the learning process based on data from individual students. AI systems assist learners to master the subject at their own pace and provide teachers with suggestions on how to help them."[3] There is also evidence that Intelligent Tutoring Systems can rival the effectiveness of human tutors[4] meaning these resources could allow all students to experience tailored support, which is seldom possible in or outside the classroom for many learners especially in schools where per capita spending is lower.[5] Furthermore, by providing responsive and appropriately challenging learning experiences, AI could increase engagement and stimulation amongst learners.[6]

These capabilities of AI could lead to beneficial structural changes. Levels of social mobility could increase if access to high quality educational opportunities became less dependent on financial means.[7] A greater breadth of intelligence could be developed if AI facilitates reforms in how learners are taught and assessed, and enables more holistic learning.

AI could also enable high quality lifelong learning for all by helping individuals access learning opportunities that are optimal for their needs[8] and by tailoring the delivery of online courses. In practice this could mean that learners of any age, in any location throughout the world, could access the learning opportunities that are right for them. Indeed, UNESCO has suggested that artificial intelligence in education could be instrumental in accomplishing Sustainable Development Goal 4[9] (Ensuring inclusive and equitable quality education and promote lifelong learning opportunities for all).

Given the promise of AI in education, an ethical approach is needed not only to guard against harmful applications of this technology, but also to avoid the underuse of resources that could impact learners positively. For instance, as the following indicates,

it is possible that students could have received a significantly more effective education during school closures due to the Covid-19 pandemic if AI systems that personalise learning had been in widespread use: “Imagine how digital personalisation systems could have helped teachers, students and parents to know what to do while studying from home when schools (and universities) closed.” (OECD Education and Skills Today, April 2020). In this case, AI techniques would have allowed systems to tailor and adapt support based on the needs of an individual learner, emulating the practice of educators to some extent. This capability makes AI well suited to supporting students in contexts where access to educators is limited.

The Institute therefore proposes the following maxim to guide ethical practice:

> Where the capabilities of AI align with the needs of learners (either directly or indirectly), the use of AI should be encouraged, providing that key ethical concerns relating to the principles of fairness, transparency, and privacy and autonomy have been addressed.

This maxim captures the dual risks of underuse and misuse of artificial intelligence in education (misuse consisting in cases where the capabilities of AI are not well suited to the needs of learners), and the need to address underlying ethical concerns associated with the general use of artificial intelligence.

Three questions arise from this maxim:

- In what contexts, and under which conditions, do the capabilities of artificial intelligence align with the needs of learners?
- How could artificial intelligence be misused in education, and what impacts could such misuses have on learners?
- In what ways could the use of artificial intelligence encroach upon a learner’s privacy and autonomy, and in what ways could the principles of fairness and transparency be violated?

With regards to question 1, wherever artificial intelligence is used in education it should be incumbent on both those responsible for the development of AI systems, and those responsible for learners' outcomes, to demonstrate how the capabilities of AI are being purposefully utilised to address the needs of learners. In section 6, methods for operationalising this principle are suggested. Sections 3 and 4 break down and explore the second two questions in further detail.

3. Avoiding the Misuse of AI in Education

AI resources that are not designed and/or used to sufficiently meet the needs of learners could result in opportunity costs, or lead to tangible harms for learners. We should be mindful that there may be instances where it is inappropriate to outsource educational processes and responsibilities to autonomous machines or algorithms, that AI resources will not always be the best tools for facilitating particular educational outcomes, and that AI resources should invariably be applied in ways that are pedagogically sound: "we should not strive for what is technically possible, but always ask ourselves what makes pedagogical sense."[10]

The following table outlines a number of potential risks presented by the misuse of artificial intelligence, offers approaches that could mitigate these risks, and poses a number of questions that aim to develop a shared understanding of what it means to misuse AI in educational contexts.

Risks	Addressing these risks	Questions for Stakeholders
The misuse of AI could erode or undermine valued skills, attributes, or aspects of learning; and narrow learners' experiences[11]. This could undermine learners' curiosity and ability to learn for themselves. Poor performance of AI systems could lead to direct harms. For instance, a system could give inappropriate mental health advice.[12] If educators are not trained and supported to use AI in an effective, augmentative way, then teachers could inadvertently become marginalised.[13] Policymakers may consider that the context of falling teacher numbers coupled with rising student numbers could justify a strategy whereby educators were increasingly replaced by AI.[14] This could have adverse consequences for learners.	Qualified educators could be involved in the design of AIEd systems.[15] Organisations developing AI could be transparent about the proxies and assumptions that are used when designing student-facing AI systems.[16] In order to protect against overuse and misuse, organisations developing AI could be explicit about the educational contexts in which the use of a particular AI system is and isn't effective.[17] There may be educational contexts where the use of AI is inappropriate, and hence discouraged.	In what contexts could the use of AI enhance learning experiences and strengthen the development of understanding, skills and attributes; and in what contexts could valued aspects of learning become marginalised if AI is misused/overused? In which educational contexts is the use of AI particularly appropriate, and in which contexts is the use of AI less appropriate, or inappropriate?

Considerations:

- If the use of AI is seen as an efficient way of reaching explicit educational outcomes, then educational outcomes/experiences that are tacitly valued may be particularly vulnerable. These might include skills such as introspection, resilience and the ability to think for oneself; or attributes such as a fondness for challenge.

- AI may be able to support a wide range of educational goals - including metacognition[18], and possibly social and emotional learning[19]. This suggests AI should not be considered inherently 'good' or 'bad' at supporting particular educational goals. Instead, attention should be paid to which AI techniques can support specific educational goals, and how.

- Note that humans will still need to be responsible for the wellbeing and performance of students, so a key consideration is when and how humans should use AI as a means of fulfilling their responsibilities.

- "...students taught through intelligent learning systems are not necessarily active decision-makers in their own learning experience. If designed with the goal of efficiency and scalability alone, intelligent learning systems not only overlook but could also strip away one of the most fundamental skills that students need to survive in the 21st century – self-actualisation" [20]

- If overused/misused, artificial intelligence could lead to a narrower, devalued learning experience.[21]

4. Addressing Further Risks from AI in Education

Fairness

Risks	Addressing these risks	Questions for Stakeholders
AI systems could exhibit biases towards different groups of learners. Biases may arise due to systems being trained on unrepresentative datasets or datasets that reflect historical biases[22], or due to the assumptions and unconscious biases of the humans developing AI systems. Due to inequalities in digital access, the increased application of AI in education could widen educational divides[23] Prevalent educational ideals in one part of the world could be imposed upon those in other parts of the world.[24]	Biases could be addressed during the development phase.[25] Due to the possibility of bias, it may be appropriate to not use AI for certain high stakes educational decisions.[26] Targeted approaches could be taken to address disparities in digital access. Organisations developing AI for education could be encouraged to be diverse and representative.	Should AI systems be benchmarked against existing levels of bias in education systems, or should they be held to a higher standard? In what high-stakes contexts, if any, should the use of AI be discouraged due to the possibility of bias? How can AI be used in education to narrow rather than widen the digital divide?

Considerations:

- It has been argued that education systems and educators are unavoidably biased, and that therefore the pragmatic criterion for AI systems should not be that they be entirely free of all bias but that their outcomes be demonstrably less biased than the system they are aiding or replacing.[27]

- It has also been argued that AI systems should be judged on how well they can resolve existing biases.[28]

Transparency/Explainability*

Risks	Addressing these risks	Questions for Stakeholders
Opaque AI systems (where reasons for actions, decisions and behaviours cannot be readily understood by humans) could mask poor performance of AI systems, which could lead to ineffective learning experiences[29] or biased outcomes.[30] Without AI systems being explainable, learners may have limited opportunities for redress; also accountability structures could be disrupted.[31] Without AI systems being explainable, learners and educators may have limited oversight of the learning process.	The opacity of AI systems could be addressed directly as part of the design process. Guidance may be needed on where and when it is acceptable to use opaque AI systems.	In which contexts, if any, should explainability be required? In which contexts, if any, is explainability less important than achieving other benefits, such as higher levels of accuracy, or increases in scale?

Considerations:

- In some cases, there may be trade-offs between levels of explainability and levels of accuracy.[32]

*An AI system is considered to be explainable if the reasons for its actions/behaviours/decisions can be understood by humans

Privacy and Autonomy

Risks	Addressing these risks	Questions for Stakeholders
Learners could be coerced and conditioned by AI systems that can manipulate their behaviours. Learners' privacy could be encroached upon. Sensitive information about students could be gathered and used as part of formal processes, e.g. admissions decisions, qualifications.	Organisations collecting, processing and sharing learners' data could be required to gain informed consent prior to data collection. Organisations collecting, processing and sharing learners' data could be required to ensure learners have oversight over how and why their data is being made use of.	In what contexts is it appropriate for AI systems to affect learners' behaviours, and in what contexts is it not appropriate? What rights should learners have over how their data is collected, processed, and shared? In what contexts should a learner's consent (or consent given on their behalf) be required? In what contexts should AI have access to learners' personal data, and in which contexts should it not? What rights should learners have over anonymised data relating to them? Should learners own their personal data, and to what extent is this possible? In what circumstances should an individual's learning data not be shared with interested third parties, such as current employers or prospective employers?

Considerations:

- Any aspect of teaching and learning may involve manipulating learners to some extent. An inspiring speech, for instance, manipulates a learner's level of motivation. That said, high levels of manipulation could encroach upon a learner's autonomy, and learners could potentially be manipulated in ways that are inherently harmful. For instance an AI system could impersonate a trusted figure and urge a learner to take harmful actions.
- It may be difficult to impose a blanket requirement of consent in practice.
- Consent may be taken as a *carte blanche* for inappropriate uses of AI.
- Data Trusts provide a model that could facilitate an individual's ownership of their data.
- In some cases, securely sharing anonymised data may benefit learners collectively. However, it has been cautioned that anonymisation of data is not fool-proof as data can be de-anonymised.[33]

5. Summary of Critical Questions

Misuse of AI in Education

- In what contexts could the use of AI enhance learning experiences and strengthen the development of understanding, skills and attributes; and in what contexts could valued aspects of learning become marginalised if AI is misused/overused?
- In which educational contexts is the use of AI particularly appropriate, and in which contexts is the use of AI less appropriate, or inappropriate?

Further Risks from AI in Education

Fairness

- Should AI systems be benchmarked against existing levels of bias in education systems, or should they be held to a higher standard?
- In what high-stakes contexts, if any, should the use of AI be discouraged due to the possibility of bias?
- How can AI be used in education to narrow rather than widen the digital divide?

Transparency/Explainability

- In which contexts, if any, should explainability be required?
- In which contexts, if any, is explainability less important than achieving other benefits, such as higher levels of accuracy, or increases in scale?

Privacy and autonomy

- In what contexts is it appropriate for AI systems to affect learners' behaviours, and in what contexts is it not appropriate?
- What rights should learners have over how their data is collected, processed, and shared?
- In what contexts should a learner's consent (or consent given on their behalf) be required?
- In what contexts should AI have access to learners' personal data, and in which contexts should it not?
- What rights should learners have over anonymised data relating to them?
- Should learners own their personal data, and to what extent is this possible?
- In what circumstances should an individual's learning data not be shared with interested third parties, such as employers?

6. Facilitating the Ethical Use of AI in Education

Whilst an ethical framework itself is intended to drive ethical decision making and therefore promote positive outcomes for learners, the Institute considers that further mechanisms may be needed in order to enable the spirit of the framework to be realised.

Based on insights from expert interviews, the Institute tentatively proposes that the following mechanisms should be implemented in order to facilitate the ethical use of AI in education. As part of the roundtable events - and the wider conversation - the Institute aims to build upon these tentative proposals. We would hence value feedback as part of this process.

Kite marks should be used to incentivise ethical practice by allowing customers (whether individuals or institutions) to clearly identify ethical products/providers. With this process, the technical features and design processes that led to a particular piece of software would be certified against a pre-established set of criteria. Drawing on points made during expert interviews[34], an aspect of the certification process could be to verify that a system has been developed via a process of participatory design, whereby stakeholders - perhaps including educators and learners - would be involved in decisions related to a product's features. This could enable AI products to better meet the needs of learners and educators. We would also suggest that kite marks be used to verify that products were developed by diverse groups of people, as this could mitigate against biases.

Coordinated efforts to educate stakeholders and develop awareness of AI in education and its ethical implications, should be used to allow individuals (including learners, educators, and those developing AI for the purposes of education) to make more informed decisions, and therefore be more discerning about how and when AI is used in educational contexts. Software developers should be trained to make ethical decisions when developing AI resources for education. Educators should be able to discern when and how AI is an appropriate tool for achieving a particular educational goal. And learners themselves should be educated about AI so that a) they can be informed participants, rather than passive subjects, where AI is being used as part of their education[35], and b) so that they are prepared to thrive in a world in which AI is becoming increasingly prevalent.[36]

Ethical training for software developers could be achieved by having compulsory ethics units as part of Computer Science degrees, or similar such qualifications. Educators could be equipped with an understanding of artificial intelligence in education as part of either initial training, continuous professional development, or both. With regards to learners, we consider that it would be appropriate for them to be educated about artificial intelligence as a core part of the curriculum. Whilst we understand that there are numerous demands on the curriculum, and correspondingly on students' time, we urge that building students' understanding of AI could benefit a) nations' economies by proactively addressing digital skills gaps, and b) individuals' life chances by ensuring they have the skills to succeed in an increasingly digital world. Indeed, in their report, Ready, Willing and Able? The House of Lords Select Committee emphasised that "all citizens have the right to be educated to

enable them to flourish mentally, emotionally and economically alongside artificial intelligence". The Institute echoes this sentiment emphatically.

The Institute also suggests that **data-ownership models** should be explored, to enable learners to have optimal levels of control over their own data.[37]

Furthermore, **guidelines for what level of evidence should be required to demonstrate the efficacy and cost-effectiveness of AI systems** (which includes the impact they have relative to other interventions/resources) will be needed. Whilst there may be cases where the benefits of an AI system can be established prior to the system's implementation, in many instances the impacts of AI on learners can only be evaluated post-implementation. For instance, it may be possible to produce evidence that an AI system can automate a particular set of tasks from pre-implementation testing. However, it is not possible to produce evidence that an Intelligent Tutoring System will have a positive impact on a population of learners until it has been implemented. A way forward could be to utilise limited implementation schemes, such as pilots or sandbox programmes, which would allow providers to amass evidence of a system's efficacy and cost-effectiveness.

7. Next Steps and Further Information

Call to action

Your perspectives matter: we want to hear from you. To have your say, please visit our website and get in touch via the contact form. We are particularly keen on hearing:

- Your views about the critical questions summarised in Section 5
- Feedback on the Institute's suggested methods for facilitating ethical practice outlined in Section 6
- Feedback on further ethical issues and/or critical questions that should be considered
- Insights that may support others' judgements on the critical questions or other ethical issues

About the Institute for Ethical AI in Education

The Institute for Ethical AI in Education is a research institution based at The University of Buckingham, and is funded by non-profit and private organisations including Microsoft Corporation and Pearson PLC.

The Institute's executive body is comprised of the Co-founders of the Institute (Sir Anthony Seldon, Priya Lakhani OBE, and Professor Rose Luckin) and the Chair of the Institute's Advisory Council (Lord Tim Clement-Jones).

The Institute's Executive Lead, Tom Moule, manages operations and research, and is the primary author of the Institute's reports.

The Institute's strategy is informed by the Advisory Council and the International Advisory Group.

For more information about the Institute, including membership of the Advisory Council and International Advisory Group, please visit our website.

https://www.buckingham.ac.uk/research-the-institute-for-ethical-ai-in-education/

References

1. https://www.mckinsey.com/industries/social-sector/our-insights/how-artificial-intelligence-will-impact-k-12-teachers
2. Luckin. R, Towards artificial intelligence-based assessment systems, 2017, Nature Human Behaviour
3. https://www.oecd-ilibrary.org/docserver/a6c90fa9-en.pdf?expires=1594115636&id=id&accname=guest&checksum=C3CAC05DFFF8EC1526043B5C93B14F0A
4. J.D, Fletcher, Effectiveness of Intelligent Tutoring Systems: A Meta-Analytical Review, 2015, Review of Educational Research
5. Seldon. A, 2018, The Fourth Education Revolution: Will Artificial Intelligence liberate or infantilise humanity, The University of Buckingham Press
6. Ibid
7. Anissa, N. ,Baker, T. , Smith, L. , (2019). Educ-AI-tion Rebooted: Exploring the future of artificial intelligence in schools and colleges. NESTA.
8. Luckin, R., Holmes, W., Griffiths, M. & Forcier, L. B. (2016). Intelligence Unleashed. An argument for AI in Education. London: Pearson.
9. Artificial Intelligence in Education: Challenges and Opportunities for Sustainable Development (2019). UNESCO. https://backend.educ.ar/refactor_resource/getBook/1097
10. Zawacki-Richter, O., Marín, V.I., Bond, M. et al. Systematic review of research on artificial intelligence applications in higher education – where are the educators?. Int J Educ Technol High Educ 16, 39 (2019). https://doi.org/10.1186/s41239-019-0171-0
11. Discussion point from meeting of The Advisory Council, July 2020
12. Insight gained from interview with Peter Westcott
13. Insight gained from interview with Tom Pieroni
14. Insight gained from interview with Lord Jim Knight
15. Proposal supported by separate insights, gained via interviews, from Dr Simon Knight (who argued in favour of participatory design approaches), Dr Alison Clark-Wilson (who argued that communities of stakeholders should be involved in critiquing AIEd resources), and Aftab Hussain (who argued that organisations developing AIEd services need to invite input from students, teachers, parents and the wider community).
16. Insight gained from interview with Dr Alison Clark-Wilson
17. Insight gained from interview with Dr Alison Clark-Wilson
18. Insight gained from interview with Dr Mutlu Cukurova
19. https://www.nesta.org.uk/blog/right-kind-ai-education/
20. Liu, Y. The Future of the Classroom, Chapter within The AI Powered State: China's approach to public sector innovation, 2020, Nesta
21. Discussion point from meeting of The Advisory Council, July 2020
22. Barton. G, Resnick. P, Turner Lee. N (2019), Algorithmic bias detection and mitigation Best practices and policies to reduce consumer harms, Brookings Institute
23. Insight gained from interview with Christian Williams
24. https://opendeved.net/2020/06/16/reflecting-on-epistemic-injustices-in-open-and-online-education/
25. https://www.mckinsey.com/featured-insights/artificial-intelligence/tackling-bias-in-artificial-intelligence-and-in-humans
26. Insight gained from interview with Martin Hamilton
27. Insight gained separately from interviews with Timo Hannay and Dr David Weinberger
28. Insight gained from interview with Jennifer Leban

29. Insight gained from interview with Dr Carmel Kent
30. Insight gained from interview with Martin Hamilton
31. Insight gained from interview with Merve Hickok
32. The Information Commissioner's Office (ICO), 2019, Project Explain: Interim Report
33. Insight gained from interview with Peter Westcott
34. See Reference #15
35. Insight gained from interview with Dr Carmel Kent
36. Insight gained from interview with Christian Williams
37. Proposal supported by separate insights, gained via interviews, from Dr Alison Clark-Wilson (who argued that learners should own their own data by right) Merve Hickok (who noted the advantages of plurality of data trusts for giving users more control and options over their data), and Lord Jim Knight (who noted the advantages of data trusts for giving users more control over their data)